Preservation in the Age of Large-Scale Digitization

A White Paper

by Oya Y. Rieger

February 2008

Council on Library and Information Resources
Washington, D.C.

ISBN 978-1-932326-29-1
CLIR Publication No. 141
Published by:

Council on Library and Information Resources
1755 Massachusetts Avenue, NW, Suite 500
Washington, DC 20036
Web site at http://www.clir.org

Additional copies are available for $20 each. Orders must be placed through CLIR's Web site.
This publication is also available online at no charge at http://www.clir.org/pubs/abstract/pub141abst.html.

The paper in this publication meets the minimum requirements of the American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data

Rieger, Oya Y.
Preservation in the age of large-scale digitization : a white paper / by Oya Y. Rieger.
p. cm. -- (CLIR publication ; no. 141)
Includes bibliographical references.
ISBN 978-1-932326-29-1 (alk. paper)
1. Library materials--Digitization. 2. Library materials--Conservation and restoration. 3. Digital preservation. I. Title.
Z701.3.D54R54 2008
025.8′4--dc22
2008002040

Contents

About the Author

Oya Rieger is interim assistant university librarian for digital library and information technologies at the Cornell University Library, where she oversees the institution's repository development, digital preservation, electronic publishing, digitization, and e-scholarship initiatives. Her responsibilities also include coordinating the library's large-scale digitization collaborations with Microsoft and Google. She is the coauthor of the award-winning *Moving Theory into Practice: Digital Imaging for Libraries and Archives* (Research Libraries Group 2000). A member of several digital imaging and preservation working groups, Ms. Rieger cochaired a group charged with developing ANSI/NISO Technical Metadata for Digital Images. Having earned a B.S. in economics, a master's degree in public administration, and an M.S. in information systems, she is currently pursuing a Ph.D. degree in a joint Cornell program with the Communication, Information Science, and Science and Technology Studies departments. Her research interests focus on the sociocultural aspects of digital technologies and scholarly communication.

Acknowledgments

I sincerely appreciate the invitation from the Council on Library and Information Resources (CLIR) to write a white paper focusing on two of my favorite topics—digitization and preservation. I am especially grateful to Kathlin Smith, CLIR's editor and director of communications, who guided me with great expertise and constant encouragement as the paper evolved from its inception to the final stages. The deep preservation background of Connie Brooks, CLIR preservation consultant, was instrumental in making sure that the paper addresses the preservation community's questions. The paper also benefited from Linda Harteker's thorough copy editing.

Special thanks go to several colleagues who were generous with their feedback during the external review. They include Bill Carney, Steve Chapman, Michele Cloonen, Paul Conway, Ricky Erway, Dale Flecker, Evelyn Frangakis, Amy Friedlander, Gary Frost, Janet Gertz, Paul Gherman, Anne Kenney, Bob Kieft, Katherine Kott, Bill Lefurgy, Anne Okerson, Vicky Reich, Brian Schottlaender, Abby Smith, and Don Waters. I also appreciate the Google Book Search, Microsoft Live Search, Million Book Project, and Open Content Alliance representatives' willingness to review the paper to confirm the accuracy of information presented about their respective initiatives. These individuals included Laura DeBonis, Jodi Healy, and Jennifer Parson from Google; Jay Girotto, Jessica Jobes, and Michel Cote from Microsoft; Denise Troll Covey and Gloriana St. Clair from Million Book Project (both from the Carnegie Mellon University Libraries); and Brewster Kahle from the Open Content Alliance.

Preface

The digitization of millions of books under programs such as Google Book Search and Microsoft Live Search Books is dramatically expanding our ability to search and find information. For scholars, it is the unparalleled scale of these undertakings that holds such promise. But it is likewise the scale of such projects that gives rise to concerns that the quality of the digitized material is inconsistent, and that the files sometimes lack important bibliographic information in their metadata.

The primary aim of large-scale digitization projects—to quickly create a critical mass of digitized books—stands in contrast to that of earlier projects, which frequently sought to create fewer, but higher-quality, scans for scholarly use. These changes in scale and quality raise a new challenge: that of maintaining the massive new collections. The point 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.

This paper examines large-scale initiatives to identify issues that will influence the availability and usability, over time, of the digital books that these projects create. As an introduction, the paper describes four key large-scale projects and their digitization strategies. Issues range from the quality of image capture to the commitment and viability of archiving institutions, as well as those institutions' willingness to collaborate. The paper also attempts to foresee the likely impacts of large-scale digitization on book collections. It offers a set of recommendations for rethinking a preservation strategy. It concludes with a plea for collaboration among cultural institutions. No single library can afford to undertake a project on the scale of Google Book Search; it can, however, collaborate with others to address the common challenges that such large projects pose.

Although this paper covers preservation administration, digital preservation, and digital imaging, it does not attempt to present a comprehensive discussion of any of these distinct specialty areas. Deliberately broad in scope, the paper is designed to be of interest to a wide range of stakeholders. These stakeholders include scholars; staff at institutions that are currently providing content for large-scale digital initiatives, are in a position to do so in the future, or are otherwise influenced by the outcomes of such projects; and leaders of foundations and government agencies that support, or have supported, large digitization projects. The paper recommends that Google and Microsoft, as well as other commercial leaders, also be brought into this conversation.

The commercial partners, as well as the participating libraries, are investing significant resources in digitization projects. How can we secure—or improve—a long-term return on this investment? Can we strike a better balance between quantity and quality? This paper outlines a range of issues relevant to the stewardship of digital resources being created by large-scale projects and to the relationship of these new resources to our print legacies. Our goal is to stimulate discussion among stakeholders and to generate productive thinking about collaborative approaches to enduring access.

CLIR is deeply grateful to Oya Rieger for so ably taking on this timely and important task. In writing this white paper, Ms. Rieger drew on her own experience and knowledge of the field as well as on responses to surveys she conducted of partners in large-scale digitization initiatives. CLIR also thanks the many experts who provided thoughtful feedback on the first draft of the paper. CLIR encourages comments from the community at large.

Charles Henry
President, CLIR

1. Introduction: Large-Scale Digitization Initiatives in the Limelight

Several research libraries are either already involved in large-scale digitization initiatives (LSDIs) or are contemplating or planning involvement in such endeavors. Two of the most visible large-scale projects, Google Book Search and Microsoft Live Search Books, have generated a flurry of debates, exchanges of opinion, and articles in various library forums and publications. Because such collaborations have far-reaching impact and are deemed inherently interesting for a general audience, the scope of commentaries has expanded to include mainstream media such as *The New Yorker* and *The Atlantic Monthly*.[1] Everyone has an opinion to express, and polarization has emerged between supporters and critics of such collaborations. There is also a group in the middle that continues to contemplate with mixed feelings the range of issues associated with LSDIs.

The goal of this white paper is to consider the potential links between large-scale digitization and long-term preservation of print and digital content, with an emphasis on research library collections. Research libraries serve as stewards of cultural heritage resources, notably books and journals, but also photographs, recordings, and other information sources. This paper focuses on books, particularly the large collections that are or may be digitized as a result of a partnership with Google, Microsoft, the Open Content Alliance (OCA), or similar agencies.

1.1 Interplay between Access and Preservation

The primary motivation of all partners in LSDIs is to make it easier to find and access books. Nonetheless, access and preservation goals are usually interrelated, since access to scholarly materials depends upon their being fit for use over time. The connection between preservation and access in the digital world is complex. For example, a library may opt to archive its digitized content as a backup in case the print counterparts are damaged or lost. However, the institution may not be able to provide online discovery and retrieval of archived digital content through a Web portal, owing to lack of funds, copyright restrictions, or other reasons.

---

[1] Jeffrey Toobin. 2007. "Google's Moon Shot: The Quest for the Universal Library." *The New Yorker* (February 5). Available at http://www.newyorker.com/reporting/2007/02/05/070205fa_fact_toobin. See also Michael Hirschorn. 2007. "The Hapless Seed." *The Atlantic Monthly* 299(5) [June]: 134-139.

While many LSDI libraries have acknowledged their intent to assume long-term responsibility for preserving digital books,[2] there is not yet a common understanding of what such responsibility entails. Who will ensure that digital content created through such initiatives remains *accessible over time*—a responsibility that is different from merely preserving it? Will responsibility for perpetual digital access be assigned to the corporate or nonprofit partners or to the libraries?

There is significant uncertainty about the long-term strategies of initiatives such as Google Book Search and Microsoft Live Search Books. These are relatively new programs and there is no evidence to suggest that the corporate and nonprofit partners have any long-term business plans for maintaining access to digitized collections or for migrating delivery platforms through future technology cycles.[3] Their online delivery and retention decisions will most likely be based on use patterns and business interests. The recent announcement that the Arts and Humanities Research Council and Joint Information Systems Committee (JISC) will cease funding the Arts and Humanities Data Service (AHDS) gives cause for concern about the long-term viability of even government-funded archiving services.[4] Such uncertainties strengthen the case for libraries taking responsibility for preservation—both from archival and access perspectives. This possibility, however, raises other questions, such as the rights to archive and provide access to digitized content still under copyright.

The interplay between the goals of access and those of preservation is also evident in discussions about the quality of digitized content resulting from current LSDI efforts. In the context of LSDIs, digital preservation can represent two distinct but related operations. It can refer to (1) preserving digital objects that result from the conversion of print materials or (2) digitizing print materials (digital reformatting) to produce digital surrogates. These two aspects of digital preservation are often conflated. The confusion arises partly from the fact that they are complementary goals and often exist within the same initiative. Although the primary incentive of the Google and Microsoft programs is to enhance access (and the image and metadata technical specifications are not pegged for digital reformatting), this does not preclude the possibility of using the resulting digital books as digital surrogates. However, some have observed that the image and optical character recognition (OCR) quality of books scanned in the LSDI projects do not adhere to reformatting best practices developed by librarians and archivists over the past 15 years. There are questions about whether materials are being converted

---

[2] See Appendix, LSDIs: Survey of Preservation Implications, question 2.

[3] According to Section 4.5 (Ownership and Control of Google Services) of the Cooperative Agreement between Google and the Committee on Institutional Cooperation (CIC), "… Google is not required to make any or all of the Google Digital Copy available through the Google Services." Available at http://www.cic.uiuc.edu/programs/CenterForLibraryInitiatives/Archive/PressRelease/LibraryDigitization/AGREEMENT.pdf.

[4] The AHDS has pioneered and encouraged awareness and use among Britain's university researchers in the arts and humanities of best practices in preserving digital data created by publicly funded research projects. The decision to cease funding is perceived as undermining the effort put into these awareness activities.

at a quality that will stand the test of time. If participating cultural institutions intend to use the resulting digital files as surrogates for analog books, or even as a just-in-case backup if an original book is lost or damaged, how can we define a digital preservation strategy that is built on the recognition that LSDIs are primarily access-driven projects?

1.2 Terminology

There is not yet a clear and consistent taxonomy for digital preservation terminology, although there are some excellent glossaries.[5] Terms such as *archiving* and *preservation* are used interchangeably, sometimes depending on the preferences of specific communities. For example, Open Archival Information System (OAIS) uses *archive* when referring to an organization that intends to preserve information for access and use by a "designated community."[6]

In this paper, *digital preservation* is used interchangeably with *archiving*. Both terms refer to a range of managed activities to support the long-term maintenance of bitstreams to make sure that digital objects are usable.[7] The definition does not include the processes required to provide continued access to digital content through various delivery methods (referred to henceforth as "enduring access" to differentiate it from bitstream preservation). 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."[8] Providing enduring access within the scope of an LSDI is a complicated responsibility. In addition to being subject to usage restrictions imposed by partners such as Google and Microsoft on digital copies provided to LSDI libraries, many digitized materials will remain in copyright for several years and cannot be made accessible online by participating libraries.

---

[5] Cornell University Library. Digital Preservation Management Tutorial. Available at http://www.library.cornell.edu/iris/tutorial/dpm/terminology/g_resources.html.

[6] The OAIS Reference Model defines a *designated community* as "an identified group of potential users of the archives' contents who should be able to understand a particular set of information." ISO 14721:2003 OAIS. Available at http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=24683&ICS1=49&ICS2=140&ICS3.

[7] The Trusted Digital Repositories report defines *digital preservation* as "the managed activities necessary for ensuring both the long-term maintenance of a bitsteam and continued accessibility of content." *Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report*. May 2002. Available at http://www.rlg.org/legacy/longterm/repositories.pdf. *Bitstream preservation* aims to keep the digital objects intact and readable. 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. Digital objects are items stored in a digital repository and in their simplest form consist of data, metadata, and an identifier.

[8] *The Preservation Management of Digital Material Handbook* is maintained by the Digital Preservation Coalition in collaboration with the National Library of Australia and the PADI Gateway. Available at http://www.dpconline.org/text/intro/definitions.html.

This paper uses the terms *mass digitization* and *large-scale digitization* interchangeably, although some draw a difference between these two terms.[9]

1.3 Outline

This paper starts with an overview of the prominent LSDIs and some of their key goals. It then provides a framework within which to assess the preservation components of digitization initiatives, including selection, content creation, and technical and organizational infrastructure. Next, the paper highlights some of the primary implications of LSDIs with regard to book collections. It concludes with a set of recommendations designed to further discussion and decision making on this important issue.

2. Overview of Leading Large-scale Digitization Initiatives

The main players in LSDIs are cultural institutions, commercial entities such as Google and Microsoft, and nonprofit groups including OCA and the Million Book Project (MBP). Although the key motivation of these stakeholders is a desire to expand access to scholarly resources, their goals differ in some ways depending on their organizational missions. The purpose of this section is to highlight the operating principles of the key players and to lay a foundation for a discussion of the preservation implications of LSDIs. Table 1 on page 9 summarizes the goals and highlights the distinguishing features of the LSDI participants.

2.1 Motivating Factors in Partnerships: Library Perspective

Some 34 cultural entities, including the 12-member Committee on Institutional Cooperation (CIC), have signed digitization agreements with Google or Microsoft. In addition, several cultural institutions are participating in the OCA and the MBP. Some libraries opt to be involved in only one initiative; others are diversifying their digitization strategies through multiple partnerships.[10]

---

[9] According to Karen Coyle, *mass digitization* is the conversion of materials on an industrial scale without making a selection of individual materials. The goal of mass digitization is not to create collections but to digitize everything, or in this case, every book ever printed. In contrast, *large-scale projects* aim to create collections and produce complete sets of documents. See Karen Coyle. 2006. "Mass Digitization of Books." *Journal of Academic Librarianship* 32(6) [November]: 641–645.

[10] Richard K. Johnson provides a useful synopsis of implications of book-digitization projects and provides examples of core library interests in digitization partnerships in his article "In Google's Broad Wake: Taking Responsibility for Shaping the Global Digital Library." *ARL: A Bimonthly Report* 250: (February 2007). Available at http://www.arl.org/bm~doc/arlbr250digprinciples.pdf.

Answers to frequently asked questions (FAQs) issued by cultural institutions participating in LSDIs indicate three major reasons for participating in large-scale projects: access, preservation, and research and development:[11]

Access. According to the FAQs, the libraries' primary motivation for partnership is to support their core mission of advancing knowledge and to transform the ways in which users search and access library content. Several participating libraries also say that these initiatives support their vision to enhance access to information in support of scholarship at local institutions and beyond. A related motivation for participation is to make the institutional collections visible worldwide.

Although most of the libraries engaged in LSDIs have significant experience in digitization, their past efforts are dwarfed by the magnitude of the Google, Microsoft, and OCA endeavors. For example, before partnering with Google, the University of Michigan, considered a leader in this domain, had been digitizing about 5,000 volumes per year. Other LSDI institutions, such as Cornell University Library or the University of Wisconsin-Madison Libraries, have created between two and three million pages of content through initiatives carried out over the past 15 years. This is an approximate equivalent of 7,000 to 10,000 book titles. At this rate, it would take them hundreds of years to convert their entire collections. Because such undertakings are costly and demanding, most libraries recognize that a logical step is to accelerate comprehensive retrospective conversion through partnerships with commercial entities.[12] Google and Microsoft have significantly raised the bar, as we are now measuring digitization initiatives in terms of millions of books, rather than millions of pages. The University of Michigan-Google LSDI is now scanning 30,000 volumes per week. At this rate, the library's entire collection (excluding materials that do not qualify) will be converted in five years.

Preservation. LSDI libraries often note the desire to ensure that library materials remain accessible to future generations as a further motivation for participation. Some institutions plan to use digitized copies as backups for works in case they go out of print, deteriorate, or are lost or damaged—to the extent allowed by copyright law. Publishers often do not keep copies of their out-of-print books, whereas

---

[11] Examples of FAQs include
Stanford: http://www-sul.stanford.edu/about_sulair/special_projects/google_sulair_project_faq.html
Harvard: http://hul.harvard.edu/hgproject/faq.html
University of Michigan: http://www.lib.umich.edu/staff/google/public/faq.pdf
Cornell: http://wiki.library.cornell.edu/wiki/x/gng.

[12] Although not at the same scale as Google and Microsoft, there are other methods to support an ambitious digitization initiative. For example, the Association of Southeastern Research Libraries, a consortium of 38 libraries, is exploring how to digitize selected portions of members' print and archival collections as a cooperative initiative. Information about this initiative is available at http://www.aserl.org/documents/ASERL_RFP_Digitization_REVISED.pdf.

libraries have a perpetual responsibility for their materials.[13]

Thirteen of 14 LSDI libraries that responded to a Web survey conducted in conjunction with this white paper expressed a commitment to archive their digitized materials (see Appendix). However, the extent of this commitment is likely to vary among institutions and has not been fully articulated.

Research and development. Some libraries, such as Stanford, perceive their participation as an opportunity to gain experience in "handling truly large amounts of digital material."[14] Some LSDI libraries mention developing advanced tools for search and retrieval and experimenting with text mining as possible activities. Grogg and Ashmore reveal that most LSDI institutions are in the early stages of exploring how to use these new digital collections and often state that "future uses are under discussion."[15]

2.2 Motivating Factors in Partnerships: Commercial Entities

Both Google and Microsoft cite the creation of a searchable database of full-text books their main motivation for partnership in LSDIs. The following sections provide an overview of the LSDIs and their access-related goals. The summaries are based on e-mail exchanges with representatives of the companies engaged in the digitization initiatives and a review of the organizations' press releases.

The summaries do not provide information on business models or financial motivations. With the exception of a few publicly available agreements, most of the contracts between the commercial partners and cultural institution are under nondisclosure clauses. RLG Programs, part of OCLC Programs and Research, is leading an effort to coordinate a series of stakeholder meetings to devise best practices in support of LSDIs. One of the outcomes of the effort is a paper by Peter B. Kaufman and Jeff Ubois on "best practices for deal-making."[16] It is based on an analysis of publicly available agreements from commercial and noncommercial mass-digitization partnerships and commentaries on these agreements and others whose documentation is not publicly available.

---

[13] See University of Michigan Library/Google Digitization Partnership FAQ. August 2005. Available at http://www.lib.umich.edu/staff/google/public/faq.pdf. University of Michigan President Mary Sue Coleman has been an outspoken advocate of the preservation role of the digital materials created through the university's partnership with Google. Noting that about five million of the books in the University of Michigan Library are either brittle or at risk because they are printed on acidic paper, she maintains that the digital copies may be the only versions of work that will survive into the future.

[14] Stanford Google Library Project FAQ. January 18, 2006. Available at http://www-sul.stanford.edu/about_sulair/special_projects/google_sulair_project_faq.html.

[15] Jill E. Grogg and Beth Ashmore. 2007. "Google Book Search Libraries and Their Digital Copies." *Searcher* (April). Available at http://www.infotoday.com/searcher/apr07/Grogg_Ashmore.shtml.

[16] Peter B. Kaufman and Jeff Ubois. 2007. "Good Terms: Improving Commercial-Noncommercial Partnerships for Mass Digitization." *D-Lib Magazine* 13 (11-12).

As of January 2008, Google Book Search participating libraries included:

Bayerische Staatsbibliothek (Bavarian State Library)
Columbia University
Cornell University
Harvard University
Ghent University Library
Indiana University
Keio University
Michigan State University
National Library of Catalonia (merged with four affiliate Catalonian libraries)
The New York Public Library
Northwestern University
Ohio State University
Oxford University
Pennsylvania State University
Princeton University
Purdue University
Stanford University
University of California
University of Chicago
University Complutense of Madrid
University of Illinois
University of Iowa
University Library of Lausanne
University of Michigan
University of Minnesota
University of Texas at Austin Library
University of Virginia
University of Wisconsin-Madison

More information about partnering libraries is available at Google Book Search Library Partners at http://books.google.com/googlebooks/partners.html.

As of January 2008, libraries participating in Microsoft's Live Search Books included:

Allen County Public Library
The American Museum of Veterinary Medicine
The British Library
Columbia University
Cornell University
The New York Public Library
Princeton Theological Seminary
University of California
University of Toronto Library
Yale University Library

2.2.1 Google[17]

The Google Book Search program aims to digitize the full text of books—both public domain and in copyright. The outcome will be a comprehensive, searchable index of a large body of published books in several languages. As of December 2007, 28 libraries were participating in the Google project, with the goal of scanning all or part of their collections and making those texts searchable online. Google is also collaborating with more than 10,000 publishers around the world in addition to its library partners. Google's business model is based on attracting as many users as possible to its site by offering a far-reaching search engine.

In 2006, a group of publishers and authors filed suit against Google, claiming that it is digitizing books without permission in order to use the information for the company's benefit. Google argues that only a limited amount of information—in the form of snippets—is displayed for materials in copyright or whose copyright status is unknown, and that this feature encourages users to obtain the book from other sources, such as bookstores and libraries. A reading of relevant publicly available documents reveals that Google's position varies on allowing participating libraries to share the digital copies of their public domain holdings with academic institutions for noncommercial purposes.

2.2.2 Microsoft[18]

Microsoft launched its Live Search Books in 2005 through a partnership with the OCA (described in section 2.3.1) to create a database of full-text books. In 2006, the company expanded its effort by recruiting additional library partners and by contracting with Kirtas Technologies[19] to undertake part of the digitization activities. Microsoft is focusing on public domain materials published before 1923. The participating libraries decide their own digitization requirements for the digital copies they will be receiving for their own use and have the option to make those copies available through the OCA in addition to through Microsoft Live Search Books. Microsoft allows academic institutions to share digital copies with other nonprofit entities as long as those entities agree not to make the files available to other commercial Internet search services.

On a complementary track, Microsoft offers the Live Search Books Publisher Program to add content through direct partnerships with publishers.[20] Live Search has distinguished itself from Google Book Search by focusing on delivering results with a unique interface and on providing advanced tools to support search and retrieval. As

---

17 Thanks to Laura DeBonis, Jennifer Parson, and Jodi Healy at Google for reviewing the information presented in this section of the paper. Additional information about the Google Book Search is available at http://books.google.com/intl/en/googlebooks/about.html.

18 Thanks to Jay Girotto, Jessica Jobes, and Michel Cote at Microsoft for reviewing the information presented in this section of the paper.

19 Kirtas Technologies: http://www.kirtas-tech.com/.

20 Microsoft Live Search Books Publisher Program: http://publisher.live.com/.

with all the search products released under the Live Search brand, Live Search Books appears as a tab on the Live Search navigation bar, along with the recently launched Live Search Academic.

2.3 Large-Scale Digitization Efforts by Nonprofit Entities

This section describes the OCA and MBP, two large, fast-moving projects by nonprofit entities with different motivations. It excludes several consortial, regional, governmental, and international initiatives as well as library partnerships with organizations such as JSTOR and Chadwyck-Healy.[21]

2.3.1 Open Content Alliance

Based on a collaboration of cultural, technology, nonprofit, and governmental organizations, the Open Content Alliance was conceived in 2005 by the Internet Archive and Yahoo![22] Its goal is to build open-access digital collections and make them available through the Internet Archive and The Open Library.[23] OCA distinguishes itself as a librarian-driven project. Unlike the Google and Microsoft initiatives, the OCA focuses on the creation of a "permanent archive" of multilingual digitized text and multimedia content.[24] All content in the OCA archive is searchable through all major search engines.[25] The files are hosted by the Internet Archive, Microsoft, and the Library of Alexandria. Other copies of these files are going into many different repository systems and may be publicly accessible from them in the future. By storing and maintaining data in multiple repositories, the OCA reports that it has been able to preserve the files, test the preservation action, and restore lost files. In addition, the images created with Microsoft funds are added to the Microsoft Live Search Books portal. Although currently focusing on public domain materials, OCA has been in discussion with some publishers to explore new business models around making copyrighted content available. OCA is partially funded by Microsoft and Adobe.

---

[21] Collaborative Digitization Programs in the United States, a Web site maintained by Ken Middleton from Middle Tennessee State University, provides links to collaborative digitization projects that focus on cultural heritage materials (http://www.mtsu.edu/~kmiddlet/stateportals.html). The June 2005 issue of *Library Hi Tech* had collaborative digitization as its theme. It is also important to acknowledge that there have been several successful regional, international, and statewide collaborations in the United States and elsewhere, although at a much smaller scale than the Google and Microsoft initiatives. For example, the Collaborative Digitization Program (http://www.cdpheritage.org/index.cfm) and the Florida Digital Archive (http://www.fcla.edu/digitalArchive/) are often cited as exemplary collaborative digitization and archiving endeavors.

[22] Open Content Alliance: http://www.opencontentalliance.org/faq.html.

[23] Internet Archive: http://www.archive.org/index.php. The Open Library: http://www.openlibrary.org/toc.htm.

[24] The OCA will seed the archive with collections from the following organizations: European Archive, Internet Archive, National Archives (UK), O'Reilly Media, Prelinger Archives, University of California, and University of Toronto.

[25] One exception to this statement is the content digitized through the Microsoft Live Books initiative and contributed to the Open Content Alliance.

Table 1. Goals and Distinguishing Features of LSDI Participants

LSDI Parties	Primary Goals	Distinguishing Feature
Research Libraries	• Support the institution's core mission of advancing knowledge • Transform the ways in which users search and access library content • Ensure that library materials remain accessible to future generations • Use digitized copies as backups • Develop advanced tools for search and retrieval; experiment with text mining	• Retain their time-tested stewardship role in collecting, organizing, managing, preserving, and providing access to information in support of learning, teaching, and research
Google Book Search	• Provide a comprehensive, searchable index of published books in several languages • Make it easier for the public to search and discover relevant books through the Google search engine • Attract as many users as possible by offering a far-reaching search engine	• Digitize the world's books to make them easier to discover
Microsoft Live Books	• Create a database of full-text books • Make it easier for the public to find relevant books • Deliver results with a unique interface and provide advanced tools to support search and retrieval	• Transform Web searches into information searches through the creation of a trusted index of authoritative content
Open Content Alliance	• Build open-access digital collections and make them available through the Internet Archive and the Open Library • Support the development of a permanent archive of multilingual digitized text and multimedia content • Preserve the files by storing and maintaining them in multiple repositories	• Create a "permanent archive" of scholarly content that can be harvested by all major search engines
Million Book Project	• Provide users with rapid, convenient access to quality resources by digitizing and making materials available on the Web • Enable equitable and worldwide access to collections to contribute to the democratization of knowledge and empowerment of a global citizenry • Maintain a test bed that stimulates and supports research in information storage and management, search engines, imaging processing, and machine translation	• Explore a range of research questions in regard to retrieval and management of large-scale and multilingual collections

2.3.2 Million Book Project[26]

The MBP is led by the Carnegie Mellon University School of Computer Science and University Libraries.[27] A distinguishing feature of MBP is its extensive digital library research agenda, which includes large-scale information storage and management, search engines for multilingual data, image processing, OCR in non-Romance languages, copyright laws and digital-rights management, and language processing. Created with a $3 million National Science Foundation

[26] In addition to the Million Book Project FAQ, information about the initiative was provided by Dean of University Libraries Gloriana St. Clair and Principal Librarian for Special Projects Denise Troll Covey at the Carnegie Mellon University Libraries.

[27] Million Book Project: http://www.library.cmu.edu/Libraries/MBP_FAQ.html.

(NSF) grant for equipment and travel, the MBP attracted international partners and matching funds exceeding US$100 million. The initial NSF-funded project officially ended in July 2007; however, partners continue to work together. Since 2001, the project has scanned more than 1.4 million books in China, India, and Egypt. It has included 26 partnering institutions, some contributing to content creation, others to the digital library research agenda. The Internet Archive is a project partner and helps acquire books for digitization. The primary countries that contribute materials for digitization (India, China, and Egypt) prefer to host the books they scan. They might eventually share their content with the Internet Archive or with OCLC, but there currently are no firm plans to do that.[28]

3. Framework for Assessing Preservation Aspects of Large-Scale Digitization Initiatives

If the library community aims to preserve the digital collections created through LSDIs, a crucial preliminary step will be to assess the community's readiness to assume such a role. Several efforts have been made in the past decade to develop standards and best practices that could provide a technical and organizational framework for managing digital preservation activities. A comprehensive discussion of this topic is beyond the scope of this paper. This section highlights some of the key components of a preservation program for digitized content.

Digital preservation within the context of LSDIs is a multifaceted topic. Two definitions are key to activity:

- *Digitizing* refers to the process of converting analog materials into digital form. If users access the digital copies instead of the analog originals (thus minimizing handling of the originals), the digital copies may be considered to have performed a preservation function. They can also perform a preservation function by serving as backups. In some cases, digital reformatting is guided by established best practices and technical specifications to ensure that the materials are being converted at a level of quality that will endure and will support future users' needs.
- *Preserving digital objects* entails the preservation of digitized materials, including those resulting from the reformatting process, to ensure their longevity and usability. In the context of this paper, the digital objects preserved may be the products of preservation reformatting or of digitization efforts in support of other purposes, such as creating a digital copy in support of online access.

The framework laid out in this section weaves through these two distinct but interrelated domains.

---

[28] This information is based on July 17, 2007, e-mail correspondence with Denise Troll Covey and Gloriana St. Clair at the Carnegie Mellon University Library.

3.1 Selection for Digitization and Preservation Reformatting

Selection and curation decisions were prominent features of early digitization initiatives. Decisions about what to digitize were influenced by traditional preservation reformatting technologies and they favored public domain materials that had enduring value for scholarship.[29] This approach was driven by a need to invest limited funds in unique aspects of institutional collections and by an interest in identifying core literature in support of research and pedagogy. Preservation of rare and brittle materials received priority; however, because early digitization technologies required that books be disbound before they could be scanned, the potential for damage to originals was a critical factor in selection decisions.

Early digitization initiatives generated lively discussion in the library community about the role of digitized content from access and preservation perspectives. Some librarians, who believed that digital files should be used as preservation master copies, expressed their commitment to preserving digital surrogates, even while acknowledging that reformatting cannot capture all characteristics inherent in the original. Some librarians believed that selection decisions should be based strictly on the need to provide access. Still other librarians felt that digital surrogates should not be considered as substitutes for the originals, but that they should still be of the highest-possible quality.

In 2004, the Association of Research Libraries (ARL) endorsed digitization as an accepted reformatting option and stated that the choice was not prescriptive and remained a local decision.[30] ARL encouraged those engaged in digital reformatting to make an organizational and financial commitment to adhere to best practices and standards.

The results of these early digitization efforts, which emphasized curatorial decisions and image quality, can be characterized as "boutique collections." Because of the magnitude of today's LSDIs, such selection criteria have largely been pushed aside. Nevertheless, the new generation of digitization projects begs the following questions in regard to selection for digitization and preservation:

1. **Should we commit to preserve all the digital materials created through the LSDIs, implement a selection process to identify what *needs* to be preserved, or assign levels of archival efforts that match use level?** According to a widely cited statistic, 20 percent of a collection accounts for 80 percent of its circulation. A multiyear OCLC study of English-language book circulation

[29] A summary of early selection approaches is provided in the following article: Janet Gertz. 1998. Selection Guidelines for Preservation. *Joint RLG and NPO Preservation Conference: Guidelines for Digital Imaging*. Available at http://www.rlg.org/preserv/joint/gertz.html.

[30] Association of Research Libraries. 2004. "Recognizing Digitization as a Preservation Reformatting Method." *ARL: Bimonthly Report* 236. Available at http://www.arl.org/bm~doc/digpres.pdf.

at two research libraries revealed that about 10 percent of books accounted for about 90 percent of circulation.[31] An analysis of circulation records for materials chosen for Cornell University Library's Microsoft initiative showed that 78 percent 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 covered by LSDIs are seldom used. Because selection for preservation can be time-consuming and expensive, the trend will likely be to preserve everything for "just-in-case" use. However, economic realities necessitate careful consideration of how much to invest in preserving unused content. This quandary is explored in Section 5.9.

2. **Will electronic access spur new demand for materials seldom used in print?** Libraries contain deep and rich collections. However, users are often hampered in locating and obtaining materials of interest because institutions use different library management systems, with varying discovery and retrieval mechanisms. Anderson argues that the 80/20 rule 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.[32] On the basis of this argument, Dempsey makes a compelling case for aggregating supply and demand at the network level rather than at the level of individual libraries.[33] Pooling the resources of many institutions' collections through LSDI partnerships, it is assumed, will find users for materials that have never been checked out.

3. **Will LSDIs' use of high-speed, automated digitizing processes disenfranchise materials needing special handling?** Most of the current LSDIs exclude special collections, which comprise rare or valuable materials including books, manuscripts, ephemera and realia, personal and professional papers, photographs, maps, fine art, audiovisual materials, and other unique documents and records.[34] Such materials require special handling because of their scarcity, age, physical condition, monetary value, or security requirements; consequently, they have high digitization costs. Early digitization efforts often included funds and services to prepare

---

[31] These data are from unpublished work by Lynn Silipigni Connaway and Edward T. O'Neill, cited in Lorcan Dempsey. 2006. "Libraries and the Long Tail: Some Thoughts about Libraries in a Network Age." *D-Lib Magazine* 12(4). Available at http://www.dlib.org/dlib/april06/dempsey/04dempsey.html.

[32] Chris Anderson. 2004. "The Long Tail." Wired 12(10). Available at http://www.wired.com/wired/archive/12.10/tail.html.

[33] Lorcan Dempsey. 2006. "Libraries and the Long Tail: Some Thoughts about Libraries in a Network Age." *D-Lib Magazine* 12(4). Available at http://www.dlib.org/dlib/april06/dempsey/04dempsey.html.

[34] There are exceptions. For instance, Cornell's partnership with Microsoft includes the digitization of in-house, special collections held at the Rare and Manuscript Collections.

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, and converting rare and special materials significantly slows the overall digitization project. Are these valuable collections being disenfranchised because of the emphasis on digitizing what can be processed quickly? Is there a danger that LSDI libraries will devote such a big share of their resources to large-scale efforts that little will remain to digitize special collections? Responding to such concerns, RLG Programs is trying to raise awareness in the special collections community of the implications of digitizing only widely held material.[35]

4. **Is there a means for recording gaps in collections and within publications?** Selection decisions for the current LSDIs are influenced by limitations of current digitization technologies as well as by the interests of the commercial partners. Improved bound-volume digitization technologies have greatly expanded the types of materials that can be digitized. However, current equipment still limits the books that can be digitized based on size (height, width, and length), condition, binding style, and paper type. The Google Five, the five pioneering libraries that signed on first with Google, report that some books are excluded because of fragile condition or problems with binding.[36] Some materials are excluded because they lack bar codes. (Although there is an established process for bar coding, some libraries avoid this process to simplify the workflow.) In addition, book sections that would require special treatment, such as maps and foldouts, often cannot be accommodated in a high-speed digitization process; such books are digitized with portions missing or not digitized at all. Incompleteness caused by missing sections has serious implications for the authenticity of digitized content. This aspect of LSDIs raises questions about tracking mechanisms for recording omitted materials and plans for adding them to the digitized corpus. Information on how LSDI survey respondents handle this issue is provided in the Appendix.

5. **How much duplication should there be in selection and digitization efforts?** In 2005, Lavoie et al. used the WorldCat union catalog to analyze book collections of the five libraries then participating in the Google Print for Libraries project.[37] After duplicate holdings across the five institutions were removed, the Google libraries together held 10.5 million unique print books out of the

---

[35] OCLC/RLG Programs, Harmonizing Digitization Program: http://www.rlg.org/en/page.php?Page_ID=21020.

[36] "The 'Google Five' Describe Progress, Challenges." 2007. *Library Journal Academic Newswire* (June). Available at http://www.libraryjournal.com/info/CA6456319.html.

[37] Brian Lavoie, Lynn Silipigni Connaway, Lorcan Dempsey. 2005. "Anatomy of Aggregate Collections: The Example of Google Print for Libraries." *D-Lib Magazine* 11(9). Available at http://www.dlib.org/dlib/september05/lavoie/09lavoie.html.

32 million in WorldCat. Of those titles, 39 percent were held by at least two of the five libraries. This suggests that four out of every ten digitized books may be redundant (assuming that digitization of titles rather than manifestations was the project goal).[38] On the basis of these preliminary data, the authors questioned the degree of redundancy associated with the digitization efforts and identified potential duplication as an area for further study.

There is a possibility for duplication both within a specific project (i.e., the same material being digitized by more than one library participating in a Microsoft initiative) and among different initiatives (i.e., the same materials being digitized both by Google and Microsoft). As these initiatives expand, the need for comprehensive collection analysis becomes more pressing. While preservation specialists acknowledge that redundancy is important for securing digital content over time, the type of redundancy that results from these approaches appears opportunistic and hence underlines the need for collections analysis across projects. It is telling that Google advertised in March 2007 for a library collections specialist to analyze the collections scanned to date and to help Google develop new library relationships with the goal of digitizing the world's books.[39]

Redundancy concerns bring registry-development efforts once again to the fore. The DLF/OCLC Registry of Digital Masters (RDM) was conceptualized in 2001 to provide a central place for libraries to search for digitally preserved materials.[40] By registering digitized objects with the RDM, a library indicates that it is committed to preserving digitized collections. One of the benefits of the registry is the assurance that one institution may not need to digitize certain materials if they are already in the registry—therefore saving resources. The potential role and current status of the registry are discussed in greater detail in Section 5.7.

6. **What legal rights do participating libraries have to preserve in-copyright content digitized through LSDIs?** Google's decision to include copyright-protected materials in its initiative has been the subject of much discussion as well as of a legal challenge. Some partners, such as the University of California, the University of Virginia, and the University of Michigan, opted to make all their collections that fit the requirements available to Google.[41] Others, including Harvard, Oxford, and Princeton Universities, decided

[38] Based on the Functional Requirements for Bibliographic Records (http://www.ifla.org/VII/s13/frbr/frbr.htm), a *work* is a distinct intellectual or artistic creation. The study included all the holdings in participating libraries and was not limited to materials selected for digitization at LSDI institutions.

[39] Ben Bunnel. 2007. Librarian Wanted: Part II. Blog posted March 22 to Google Librarian Central. Available at http://librariancentral.blogspot.com/2007/03/librarian-wanted-part-ii.html.

[40] DLF/OCLC Registry of Digital Masters: http://www.oclc.org/digitalpreservation/why/digitalregistry/. The registry database is available at http://purl.oclc.org/DLF/collections.

[41] Information about the decision of libraries to contribute both public domain and in-copyright material is obtained from their press releases and aforementioned FAQs.

to limit their participation to public domain content. An analysis by Brian Lavoie of OCLC found that 80 percent of the original five Google libraries' materials were still in copyright.[42] This aspect of the LSDIs raises the issue of the participating libraries' legal rights to preserve copyrighted content digitized through LSDIs. For example, is it legally permissible for a library to rescan originals that are not in the public domain to replace unusable or corrupted digital objects? What are the copyright implications of migrating a digital version of materials in copyright from TIFF to JPEG2000 file format?

Section 108 of the U.S. Copyright Law articulates the rights to and limitations on reproduction by libraries and archives;[43] however, the right to take action to preserve digitized content that is copyright protected is still under study by the Section 108 Study Group convened by the Library of Congress.[44] The Study Group is charged with updating the Copyright Act's balance between the rights of creators and copyright owners and the needs of libraries and archives within the digital realm. The group is also reexamining the exceptions and limitations applicable to digital preservation activities of libraries and archives.[45]

When the CIC libraries joined the Google Initiative, they decided to archive only materials in the public domain and opted not to receive digital copies of materials in copyright until a general preservation exception (the right to preserve materials in copyright) is added to Section 108.[46]

3.2 Content Creation

Digital preservation requires a sequence of decisions and actions that begin early in an information object's life cycle. Standard policies and operating principles for digital content creation are the foundation of a successful preservation program. Table 2 summarizes the

---

[42] Brian Lavoie, Lynn Silipigni Connaway, Lorcan Dempsey. 2005. "Anatomy of Aggregate Collections: The Example of Google Print for Libraries." *D-Lib Magazine* 11(9). Available at http://www.dlib.org/dlib/september05/lavoie/09lavoie.html.

[43] Section 108 provides exceptions that allow libraries and archives to undertake certain activities, otherwise not permitted, for purposes of preservation and, in some cases, replacement. There is intense discussion about the elements of Section 108 in regard to who can do what for what purposes and under what conditions. Circular 92: http://www.copyright.gov/title17/92chap1.html#108.

[44] Section 108 Study Group: http://www.loc.gov/section108/.

[45] Peter Hirtle analyzes a range of preservation-related issues and concludes that copyright in regard to digital preservation is an especially murky area because it is necessary to copy digital information in order to preserve it. See http://fairuse.stanford.edu/commentary_and_analysis/2003_11_hirtle.html.

[46] Section 4.11 (Release of In-Copyright Works Held in Escrow) of Google's agreement with CIC states that Google will hold the "University Copy" of scanned works. It lists the conditions under which the copies will be released to the contributing libraries, including if the in-copyright work becomes public domain or if the library party has obtained permission through contractual agreements with copyright holders. See http://www.cic.uiuc.edu/programs/CenterForLibraryInitiatives/Archive/PressRelease/LibraryDigitization/AGREEMENT.pdf.

activities involved in creating digital collections. These guidelines apply to digitization initiatives with preservation as an explicit mandate. Most of the current LSDIs do not fall within that category.

The purpose of this section is to inform discussions of the differences between access- and preservation-driven reformatting. It reviews the following aspects of content creation, all of which are critical in producing high-quality collections:

- technical specifications for image-quality parameters for master and archival files
- requisite preservation metadata with descriptive, administrative, structural, and technical information to enhance access as appropriate, enable content management, and facilitate discovery and interoperability
- quality-control protocols for digital images and associated data

Table 3 provides examples of digitization specifications used in different initiatives for digital copies received by libraries, including resolution, bit depth, image format, capture device information, and metadata standards supported. Additional content-creation information from sample LSDI libraries, including quality control parameters, is provided in the Appendix.

Table 2. Framework for a Digitization Project

Creating digital collections encompasses a diverse array of activities. The list of main functional areas that follows represents a wide range of skills:

Selection
- material selection based on research, learning, and teaching needs
- copyright-status assessment

Requirements analysis to set technical requirements for:
- digitization
- metadata
- access and use
- other repurposing areas (e.g., print on demand)

Preparation
- conservation, disbinding, tagging
- physical volume organization by content or format type

Digitization
- digitization (in-house or outsourced)
- image processing
- creation of archival and derivative files
- structuring

Quality control
- development of a QC strategy
- selection of QC tools
- development of assessment workflow
- plan for correcting and reintegrating unacceptable images and other deliverables

Metadata
- descriptive, structural, administrative, preservation
- controlled vocabulary, taxonomies, ontologies
- selecting and implementing standards for interoperability, discovery, etc.
- file-naming conventions and persistent IDs
- OCR

Technical development
- repository and storage plan
- digital content delivery platform (image database)
- discovery and navigation tools
- Web services
- Web design and development

Project management
- workflow coordination
- financial management
- assessment and usability analysis
- promotion
- user support

Life cycle management
- preservation strategies and procedures
- ongoing content, metadata, application revisions, additions, etc.

Table 3: Examples of Digitization Specifications Used in Different Initiatives for Digital Copies Received by Libraries

	Resolution/ Bit Depth	Image Format	Capture Device Information	Metadata
University of Michigan Library **Digital Content Received from Google**	Most pages 600 dpi, 1 bit * Information about whether these are 600-dpi native images or interpolations based on lower resolutions is not disclosed	TIFF ITU G4 compression Pages with illustrations provided in 300 dpi JPEG2000	Not available	MARC is used for descriptive metadata; technical and preservation metadata recorded in local text format; METS[47] profile under development
Cornell University Library **Digital Content Received from Microsoft (Kirtas)**	300–400 dpi 8–24 bit	JPEG Considering switching to JPEG2000	APT 2400 or Scribe for OCA partners 100% quality control by Kirtas	METS schema is used for recording MARC and Z39.50 (JHOVE-extracted technical metadata is recorded in MIX format) and the mandatory fields of PREMIS metadata
Open Content Alliance	400–600 dpi, 12 bit	JPEG2000	Scribe[48]	Bibliographic metadata in MARC binary, MARC XML, and Dublin Core; technical metadata maintained in XML format
Million Book Project	Predominantly 600 dpi, 1 bit	TIFF, ITU Group 4	Varies, including Minolta PS7000 scanner	No central database or requirements; practice varies among participating libraries in Egypt, China, and India

[47] The Metadata Encoding and Transmission Standard (METS) schema is used for encoding descriptive, administrative, and structural metadata for digitized pages. More information about METS is available at http://www.loc.gov/standards/mets/. MIX is an XML schema developed for recording and managing the Technical Metadata for Digital Still Images (ANSI/NISO Z39.87-2006). The standards page is at http://www.loc.gov/standards/mix/.

[48] The Scribe system used by the OCA requires a manual operator to turn pages and monitor the images. OCA Scribe includes two commercial-grade cameras, and the open books rest at 90° on an adjustable spring cradle.

3.2.1 Image-Quality Procedures for Large-Scale Digitization Initiatives

LSDIs use a variety of 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 and whether current capture quality will support future viewing and processing needs.[49] Because of nondisclosure agreements and varying practices based on proprietary digitization configurations, it is hard to obtain information on LSDIs' current digitization and metadata specifications. Some examples gathered as a result of the LSDI preservation survey are provided in the Appendix. Most of the participating libraries have been involved in digitization initiatives for well over a decade, so it is useful to start with a synopsis of prevailing preservation standards and best practices for digital material.[50]

The term *digital imaging* entered the library lexicon predominantly through the preservation community, which had an early interest in using digitization as a reformatting tool. Most of the initial projects approached the image-quality issue from a perspective that was heavily influenced by microfilming requirements and micrographic industry standards. The recommendations that came out of 1990s image-benchmarking studies still influence discussions about image quality. These requirements were developed to ensure that a book would need to be digitized only once. The goal was to create "preservation-worthy" images that would be faithful reproductions of the original material and rich enough in quality to justify investing in their long-term archiving. There was also an emphasis on using digitization techniques that would minimize damage to original materials. Although user needs were reflected in the benchmarks, most of the effort was directed at capturing as much detail and nuance from the print materials as possible. A difference was drawn between master (archival) images, which are optimized for longevity and repurposing, and derivative (access) files, which support specific uses, such as printing or online viewing. There was strong endorse-

[49] David Bearman offers an opinion piece on Jean-Noël Jeanneney's comments on the Google initiative. One of the critiques presented by Jeanneney is Google's sloppy imaging of books because its primary interest is harvesting works to link to advertising. See David Bearman. 2006. "Jean-Noël Jeanneney's Critique of Google: Private Sector Book Digitization and Digital Library Policy." *D-Lib Magazine*12(12). Available at http://www.dlib.org. On the basis of a case study, Paul Duguid has pointed out problems with scans, metadata, and edition information in Google Book Search. He concludes that Google's powerful search tools cannot make up for a lack of metadata. See Paul Duguid. 2007. "Inheritance and Loss? A Brief Survey of Google Books." *First Monday* 12(8) [August]. Available at http://www.firstmonday.org/issues/issue12_8/duguid/. The discussion as a reaction to the article is available at http://radar.oreilly.com/archives/2007/08/the_google_exch.html.

[50] For a thorough review of the digitization approaches used during the past 10 years, see Steven Puglia and Erin Rhodes. 2007. "Digital Imaging: How Far Have We Come and What Still Needs to be Done?" *RLG DigiNews* 11(1) [April 15]. Available at http://www.rlg.org/en/page.php?Page_ID=21033.

ment for using lossless compression for the master images,[51] and TIFF became the de facto archival format.

The digitization efforts of the early 1990s resulted in well-endorsed best practices and benchmarks that have been widely adopted. However, today's digitization efforts challenge some of the prevailing practices not only because of their scale but also because of the transformation of digital library technologies and user preferences.

- *Doing Today's Job with Yesterday's Tools*
  Imaging devices have improved since the mid-1990s; however, we continue to rely on the results of early work on assessing digitization devices and image quality. Early efforts emphasized fixed spatial resolution and bit depth, a quantitative approach that does not always indicate the capture quality of digitization equipment. The same resolution setting, such as 600 dpi, in two machines may render different results. As Puglia and Rhodes point out, the focus has shifted to high spatial resolution and high bit sampling.[52] The trend is moving from testing the capabilities of digitization equipment to assessing specific device performance parameters. High spatial resolution and bit depth are ideal; however, they alone do not guarantee satisfactory images. Therefore, more emphasis must be placed on assessing outcomes.

- *Role of Image-Quality Targets*
  To verify the calibration of the scanning equipment and to ensure the best-possible images, early initiatives used image-quality targets. Libraries required delivery of specified scanned technical targets during the installation and configuration of scanning equipment, and they relied on these targets during the production of images to assess resolution, tonality, dynamic range, noise, and color. These targets were also seen as instrumental in preserving technical information that may be needed for certain future preservation actions, such as file migration.[53]

  Today, LSDIs do not consistently use such quality targets. This may lead to a lack of common protocols in assessing image quality and to making adjustments (e.g., changing color space). The Digital Image Conformance Evaluation (DICE) tool being developed at the Library of Congress by Don Williams, Peter Burns, and Michael Stelmach is promising and will result in an assessment target and associated software for automated analysis.[54]

---

[51] For a discussion on the difference between lossy and lossless compression, see Moving Theory into Practice: Digital Imaging Tutorial, Cornell University Library. Available at http://www.library.cornell.edu/preservation/tutorial/intro/intro-07.html.

[52] Puglia and Rhodes 2007 op cit.

[53] For example, Macbeth ColorChecker is used to inspect color fidelity and to control color space during file format migration and other image processing activities.

[54] The information about DICE is based on e-mail correspondence with Don Williams, technical imaging consultant, standards and image quality, in July 2007.

- *File Format and Compression*
  TIFF has been the de facto file format for archival copies of digital images since the early 1990s.[55] Adobe Systems controls the TIFF specification, which has not had a major update since 1992.[56] TIFF has many advantages, such as support of lossless compression, which is strongly favored by the preservation community because it retains full pixel information. However, some institutions engaged in large-scale efforts are considering a switch to JPEG2000, which can be lossy or lossless depending on the compression algorithm used.[57] It is an International Standards Organization (ISO) standard and permits a wide range of uses. It allows metadata to be built into the file. Other advantages include scalability by resolution (several resolution levels are included in one file to support different views), availability of color channels to manage color appearance information, and bit-depth support up to 48 bits.[58] JPEG2000 uses a compression technique based on wavelet technology, which produces smaller file sizes that are more efficient to store, process, and transfer than large files are. However, the standard is not yet commonly used and there is not sufficient support for it by Web browsers. The number of tools available for JPEG2000 is limited but continues to grow.

3.2.2 Preservation Metadata

PREservation Metadata: Implementation Strategies (PREMIS) defines *preservation metadata* as "the information a repository uses to support the digital preservation process."[59] It includes data to support maintaining viability, renderability, understandability, authenticity, and identity in a preservation context. Although the theory and standards behind preservation metadata are sound, its long-term cost-effectiveness and utility remain unknown.

Preservation metadata incorporates a number of metadata categories, including descriptive, administrative (including rights and permissions), technical, and structural. PREMIS emphasizes recording digital provenance (the history of an object). Documenting the attributes of digitized materials in a consistent way makes it possible

---

[55] The 2006 report *Digital Image Archiving Study*, issued by the Arts and Humanities Data Service, includes a comprehensive discussion of various raster image file formats and reviews their advantages and disadvantages for preservation purposes. Available at http://ahds.ac.uk/about/projects/archiving-studies/digital-images-archiving-study.pdf.

The March 2007 CENDI report assessed file formats for preserving government information. It is not focused on digital images; however, it is a useful document for understanding format assessment factors for digital preservation. CENDI Digital Preservation Task Group. March 2007. *Formats for Digital Preservation: A Review of Alternatives*. Available at http://www.cendi.gov/publications/CENDI_PresFormats_WhitePaper_03092007.pdf.

[56] TIFF: http://partners.adobe.com/public/developer/tiff/index.html.

[57] JPEG2000 file format information: http://www.jpeg.org/jpeg2000/.

[58] For an in-depth discussion of file formats, see Tim Vitale. Digital Image File Formats—TIFF, JPEG, JPEG2000, RAW, and DNG. July 2007, Version 20. Available at http://aic.stanford.edu/sg/emg/library/pdf/vitale/2007-07-vitale-digital_image_file_formats.pdf.

[59] PREMIS: http://www.oclc.org/research/projects/pmwg/.

to identify the provenance of an item as well as the terms and conditions that govern its distribution and use. In digitization initiatives with homogeneous and consistent practices, it may be sufficient to capture preservation metadata at the collection level without recording details at page level.

Although incorporated in preservation metadata, technical metadata merits special mention because of its role in supporting preservation actions. Published in 2006, ANSI/NISO Z39.87 Technical Metadata for Still Images lays out a set of metadata elements to facilitate interoperability among systems, services, and software as well as to support continuing access to and long-term management of digital image collections.[60] It includes information about basic image parameters, image quality, and the history of change in document processes applied to image data over the life cycle. The strength and weakness of Z39.87 is its comprehensive nature. Although in many ways an ideal framework, it is complex and expensive to implement, especially at the image level. While most of the technical metadata can be extracted from the image file itself, some data elements relating to image production are not inherent in the file and need to be added to the preservation metadata record.[61] Google does not allow access to its digitization centers because of the proprietary hardware and software in use. Therefore, it may not be possible to gather certain technical specifications for image production in its LSDI. The role of technical metadata (or lack thereof) in facilitating preservation activities is not yet well documented.

3.2.3 Descriptive and Structural Metadata

It is difficult to consider an image to be of high quality unless there is requisite metadata to support identification, access, discovery, and management of digital objects.[62] Descriptive metadata ensures that users can easily locate, retrieve, and authenticate collections. The current LSDIs rely on bibliographic records extracted from local Online Public Access Catalogs (OPACs) for descriptive metadata. Compared with early digitization initiatives, minimal structural metadata are captured. There is an effort to promote the use of persistent IDs both by search engines and by participating libraries to ensure that globally unique IDs are assigned to digitized books.

Structural metadata facilitates navigation and presentation of digital materials. It provides information about the internal structure

---

[60] Z39.87: Data Dictionary—Technical Metadata for Digital Still Images. Available at http://www.niso.org/standards/standard_detail.cfm?std_id=731.

[61] Metadata-extraction tools such as JHOVE and NLNZ Metadata Extractor Tool generate standardized metadata that is compliant with PREMIS and Z39.87.

[62] The NARA Technical Guidelines for Digitizing Archival Materials for Electronic Access define approaches for creating digital surrogates for facilitating access and reproduction; they are not considered appropriate for preservation reformatting to create surrogates that will replace original records. See Steven Puglia, Jeffrey Reed, and Erin Rhodes. June 2004. Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files Raster Images. Available at http://www.archives.gov/preservation/technical/guidelines.pdf.

of resources, including page, section, chapter numbering, indexes, and table of contents. It also describes relationships among materials and binds the related files and scripts through file naming and organizing files in system directories. Current LSDI digitization processes often do not capture structuring tags such as title page, table of contents, chapters, parts, errata, and index. Gathering and recording such data are usually neither feasible nor cost-effective within an LSDI workflow. It is important to include structural metadata in the definition and assessment of digital object quality. For example, checking the availability of structural metadata for complex materials such as multivolume books is critical to retaining the relationship information among multiple volumes.

3.2.4 Quality Control

Quality control (QC) is an essential component of library digitization initiatives.[63] It includes procedures and techniques to verify the quality, accuracy, and consistency of digital products encompassing images, OCR output, and other metadata files. The key factors in image-quality assessment are resolution, color and tone, and overall appearance. Intent, such as reproducing a physical item, restoring to original appearance (e.g., removing stains), improving legibility, or optimizing for Web presentation or printing, is also important.

Sometimes a distinction is drawn between quality control and quality review (or quality assurance). The former refers to the vendor or in-house inspection conducted during production; the latter indicates the inspection of final products by project staff. (In this paper, QC will be used to refer to both processes.) Implementing a QC program can be very time- and labor-intensive, and requires special skills and equipment. Although automated tools[64] exist for inspecting certain aspects of quality (such as file naming and integrity checks), some quality elements, such as missing pages and imaging distortions, can be detected only through visual inspection.

The initial QC efforts of the library community were quite thorough and often involved 100 percent QC, with visual digital and print page comparison looking for subtle indicators such as wavy patterns, bandings, and Newton's rings. Today, there are well-established image-quality assessment processes; however, they are based on digitizing small subsets of library materials. Although there is some reliance on automated QC tools, most quality assurance is done manually. 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 need to be reevaluated to decide what best suits the budget, technical infrastructure, staff qualifications, materials, and project time line.

---

[63] Oya Y. Rieger. 2000. "Establishing a Quality Control Program." Pp. 61-83 in *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, by Anne R. Kenney and Oya Y. Rieger. Mountain View, Calif.: Research Libraries Group.

[64] For example, Cornell University Library uses a locally developed application to automatically check images matching MD5 checksums, availability of OCR and position data, detection of blank pages, etc.

Currently, the Google initiative is not correcting images based on QA procedures conducted by the participating libraries. However, it gathers library-partner feedback on image condition in correlation with its own analysis to create automated QC methods to improve image quality. As indicated in the survey results in the Appendix, the participating libraries are mainly recording trends and patterns, for the purpose of improving the quality of future scans. Microsoft and the OCA have workflows in place for correcting images based on the feedback received from participating libraries.

The Google initiative often has been criticized for producing scans with missing pages or poor image quality (e.g., blurs or other markings). Townsend argues that a closer look at the digitized materials on the Google Books site raises concerns about the variability of image quality and erroneous or incomplete metadata, especially in serials literature.[65] He is concerned that these problems will compound over time and that it may be difficult to go back and make corrections when the imperative is to move forward. Errors and omissions may become an important issue, especially if there is full reliance on digital copies and print versions are not readily available. Of course, even if the quality is inconsistent, the digitized books support discovery and can provide some level of emergency backup if something were to happen to the print books. However, inconsistent quality and gaps pose a serious preservation issue if the digitized books are used as an excuse to discard all the original books.

The digitization process captures the page image, and OCR tools are required to extract the text in a machine-readable form. The term *OCR* generally refers to the process by which scanned images are electronically "read" to convert them into text to support full-text searching and other processes that require editable text. Although OCR used to be optional (often because of funding constraints) in digitization projects, today's LSDIs automatically include such a process in order to create sophisticated full-text indexes to enable retrieval of materials by keyword.

The accuracy of OCR depends on the quality of images and the capabilities of OCR engines in processing different font types, languages, and such. This is particularly true for older books with ancient and pale fonts. Obtaining accuracy close to 100 percent usually requires some level of manual correction. Human intervention in large-scale efforts is minimal, so OCR files do not typically go through a quality control process to identify errors. Some experts believe that the 98-99 percent accuracy achieved by automated OCR is good enough to meet indexing and discovery needs. Other scholars, such as Jean-Claude Guedon, have expressed concern that centuries of progress toward increasingly accurate and high-quality printing

---

[65] Robert B. Townsend. 2007. "Google Books: What's Not to Like?" *AHA Today* (April 30). Available at http://blog.historians.org/articles/204/google-books-whats-not-to-like. Also see footnote 50.

could be reversed owing to lack of OCR quality control and high standards for LSDIs.[66]

3.3 Technical Infrastructure

Numerous factors put digital data at risk. Many technologies disappear as product lines are replaced, and backward compatibility is not always guaranteed. The vulnerable elements of the technical infrastructure for digital image collections include the following:

- storage media: at risk for mishandling, improper storage, data corruption, physical damage, or obsolescence
- file formats and compression schemes: at risk for obsolescence
- various application, Internet protocol, and standard dependencies: at risk for impact of updates and revisions on dependent processes and operations

The sidebar at left lists some of the technical procedures involved in preserving digital content. The curatorial strategies listed are being addressed in various library forums. Digital preservation infrastructure relies on a robust computing and networking infrastructure and a scalable storage strategy. This section primarily addresses storage issues, which pose a major challenge for data management and storage architectures because of the sheer amount of data presented by LSDIs.

E-science data initiatives have introduced libraries to the challenges associated with large-scale database storage and retrieval.[67] Nonetheless, many participating libraries still have limited experience in data management. An extensive review by the British Library revealed that storage technologies continue to evolve and that data-storage vendors are coming up with new standards and solutions.[68] This trend requires the development of solutions that can accommodate expanding content and emerging storage technologies. The findings of the recent Getty Research Institute survey also pointed out the need to rethink infrastructure and storage models.[69]

Selected Technical Features of a Program in Support of Bit Stream Preservation

- creation of a repository model to ingest, monitor, manage, and archive digital objects and associated metadata, files, and scripts
- development and implementation of an ingest workflow and quality control measures to verify authenticity and completeness of ingested content
- creation and management of preservation metadata (including technical metadata)
- identification of properties to preserve in digital objects
- continuous monitoring and management of digital content to detect bit corruption, loss, or obsolescence
- record of provenance and change history for all objects
- programs in support of various preservation strategies, including refreshing, migration, replication, normalization, and emulation (both for preventive measures and for staying abreast of standards and technologies)
- disaster-prevention, recovery, and contingency plans
- periodic review and updating of preservation procedures
- mechanisms for monitoring triggers for preservation action (e.g., file format migration, file corruption)
- security measures
- technical audits

---

[66] U.S. National Commission on Libraries and Information Science. *Mass Digitization: Implications for Information Policy*. Report from Scholarship and Libraries in Transition: A Dialogue about the Impacts of Mass Digitization Projects. Symposium held March 10-11, 2006, at the University of Michigan, Ann Arbor. Available at http://www.nclis.gov/digitization/MassDigitizationSymposium-Report.pdf.

[67] The following article covers the storage and networking challenges faced by educational institutions in supporting the emerging technical infrastructures: Thomas J. Hacker and Bradley C. Wheeler. 2007. Making Research Cyberinfrastructure a Strategic Choice. *Educause Quarterly* 30(1): 21-29. Available at http://www.educause.edu/ir/library/pdf/EQM0713.pdf.

[68] Jim Linden, Sean Martin, Richard Masters, and Roderic Parker. February 2005. *The Large-Scale Archival Storage of Digital Objects*. Digital Preservation Coalition Technology Watch Series Report 04-03. Available at http://www.dpconline.org/docs/dpctw04-03.pdf.

[69] The International Digital Preservation Systems Survey conducted by Karim Boughida and Sally Hubbard from the Getty Research Institute intended to provide an overview of digital preservation system implementation. Comments are based on e-mail exchange with Boughida in June 2007. In the context of the survey, *a digital preservation system* is defined as an assembly of computer hardware, software, and policies equivalent to a trusted digital repository with a mission of providing reliable, long-term access to managed digital resources.

According to the Getty international survey of digital preservation systems, 66 percent of the 316 institutions surveyed had less than 10 terabytes (Tb) of data, and their storage costs ranged from $400 to $15,000 per Tb. The wide range reflects the lack of common metrics in projecting and reporting such expenses. With the exception of two sites that store video, Digital Library Federation (DLF) members representing many LSDIs that responded to the survey reported storing data in the range of 2–20 Tb. Cornell University Library's copy of each digital book created through the Cornell-Microsoft partnership is approximately 700 megabytes (Mb). Based on this per-book estimate, Cornell anticipates having to store approximately 60 Tb of digital content, representing nearly 100,000 volumes, during the first year of the initiative. By comparison, Cornell accumulated only 5 Tb of content through 15 years of digital imaging activities.

Currently there is neither a metric nor a methodology for estimating resources required for storage. Moreover, storage expenses depend on local information technology (IT) and repository infrastructures and configuration, making generalizations difficult. As libraries acquire more and more digital content, it will be important to understand the current and projected costs of storage.[70] While the storage hardware costs can be obtained from vendors, there is little detailed information on the operational costs associated with storage. Factors that influence overall costs include quantity of data, server configuration, storage media, storage-management software, projected data-storage needs, data-access time, data-transfer rate, access services supported, and redundancy and backup protocols. Life Cycle Information for E-Literature (LIFE), a JISC-funded joint venture, developed a methodology to calculate the long-term costs and future requirements of preserving digital assets.[71] Project staff found that it costs £19 (about US$38) to store and preserve an e-monograph in Year 1; by the tenth year, the total life cycle cost is £30 (US$51). These costs include acquisition, ingest, basic metadata, access, storage, and preservation, but do not include creation. It is important to be cautious about generalizing the LIFE estimates as they are based on a small file of approximately 1.6 Mb using specific workflow and process, as compared with the estimated 700 Mb per digital book created through the Cornell-Microsoft partnership.

An assessment by the National Archives of Sweden revealed that storage media represent only 5 percent to 10 percent of total storage expenses; the bulk of costs are associated with hardware, software, support, maintenance, and administration.[72] Some libraries, such as Cornell University Library, pay their home institutions for band-

---

[70] Moore et al. describe current estimates of both disk and tape storage based on operational experience at the San Diego Supercomputer Center, which operates a large-scale storage infrastructure. See Richard L. Moore, Jim D'Aoust, Robert McDonald, and David Minor. 2007. "Disk and Tape Storage Cost Models." *Archiving* (May): 29-32.

[71] The LIFE Project: Lifecycle Information for E-Literature. 2006. Available at http://www.ucl.ac.uk/ls/life/.

[72] Jonas Palm. *The Digital Black Hole*. 2006. Stockholm: Riksarkivet. Available at http://www.tape-online.net/docs/Palm_Black_Hole.pdf.

width consumed (network usage-based billing), adding yet another storage-cost element. The bandwidth charge for transferring a digital book online from Victor, New York (Kirtas), to Cornell is about 95 cents.[73] Scaling this estimate to 100,000 books results in an anticipated additional project cost of nearly $95,000.[74]

Storage requirements need to be assessed from an information life cycle management perspective to ensure a sustainable storage strategy that balances costs, data management strategies, preservation priorities, and changing use patterns. The life cycle approach requires more complex criteria for storage management than do automated storage procedures, such as in hierarchical storage management.[75] There is also an increasing emphasis on the virtue of distributed preservation services, such as replicating content at different locations.

3.4 Organizational Infrastructure

Technology alone cannot solve preservation problems. 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. Clareson illustrates the gap between digitization and digital preservation practice by pointing out that "except for inclusion in rights and licensing policies, digital holdings are not included in the majority of policy statements for many areas of institutional operation, from mission and goals to emergency preparedness, to exhibit policies."[76] The challenge is not only to incorporate the preservation mandate in organizational mission 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 archiving in support of short-term goals. There are also significant differences between a preservation program that focuses on bitstream preservation and one that encompasses the processes required to provide enduring access to digital content.

---

[73] The average size of a digitized book is about 700 megabytes.

[74] As an alternative to direct network transfers, the storage team has explored the methods and costs involved in shipping external hard drives or high-storage computers loaded with page image files between Victor and Ithaca. Using physical storage devices for transportation is a cumbersome, risky, and expensive process. In addition to insecurities of physical transport, there are hardware incompatibilities between the Microsoft and Sun platforms.

[75] Hierarchical storage management is a data-storage method to ensure cost-effectiveness based on data-usage patterns. The system monitors how data are used and automatically moves data between high-cost (faster devices) and low-cost (slower devices) storage media.

[76] Tom Clareson. 2006. "NEDCC Survey and Colloquium Explore Digitization and Digital Preservation Policies and Practices." *RLG DigiNews* 10(1) [February]. Available at http://www.rlg.org/en/page.php?Page_ID=20894.

Key Organizational-Infrastructure Requirements for Preservation Programs

- description and characterization of the preservation mandate in the organizational mission statement and inclusion of supporting programs in strategic planning
- identification of the scope and extent of preservation activities and priorities (with the recognition that it is not possible to preserve everything)
- resources allocated in a way that indicates an institutional commitment to ensuring the integrity, authenticity, and usability of digital content
- technical requirements and best practices for digital content creation
- compliance with community standards and best practices for digital preservation, access, and interoperability
- cost projections and analysis
- financial planning and management by considering in-house and outsourcing options from the perspectives of cost-effectiveness and operational efficiency
- identification of staff skills and staffing patterns required to implement preservation strategies
- ongoing training and professional development opportunities for staff
- policies for selecting, reselecting, and deselecting content for preservation
- plans for moving digitization projects supported by grant funds or special allocations into mainstream programs
- procedures to meet archival requirements pertaining to provenance, chain of custody, authenticity, and integrity
- policies and documentation procedures in support of intellectual property rights
- emergency-preparedness and disaster-recovery plans
- identification of collaboration and cooperation opportunities at the local, national, and international levels
- technology forecasting on trends in digital preservation
- assessment of risks by monitoring technological changes
- consistent and documented policies, procedures, and practices for the overall preservation program

Organizational preservation infrastructure—mandate, governance, and funding models—is emerging as a critical factor in determining success. The box above lists the organizational infrastructure requirements needed to support preservation programs. There are several formal standards and best practices in place.[77] The following are some examples:

Open Archival Information System (OAIS). The OAIS reference model addresses a full range of preservation functions, including ingest, archival storage, data management, access, and dissemina-

[77] A review of prevailing preservation standards and protocols is discussed in Nancy Y. McGovern. 2007. "A Digital Decade: Where Have We Been and Where Are We Going in Digital Preservation?" *RLG DigiNews* 11(1) [April 15]. Available at http://www.rlg.org/en/page.php?Page_ID=21033#article3.

tion.[78] Specifically applicable to organizations with long-term preservation responsibilities, it has provided a framework and a common language for digital preservation discussions and planning activities, especially for their technical and architectural aspects.

Trustworthy Repositories Audit & Certification (TRAC). An OCLC/RLG Programs and National Archives and Records Administration (NARA) task force developed the Audit Checklist for Certifying Digital Repositories as a tool to assess reliability, commitment, and readiness of institutions to assume long-term preservation responsibilities.[79] With the revision and publication of the tool as the TRAC checklist in March 2007, the Center for Research Libraries expressed its intention to contribute to related digital repository audit and certification, including guiding further international efforts on auditing and certifying repositories.[80]

Digital Repository Audit Method Based on Risk Assessment (DRAMBORA). Released in March 2007 for public testing and comment, the DRAMBORA toolkit aims to facilitate internal audit by providing preservation administrators with a means to assess their capabilities, identify their weaknesses, and recognize their strengths.[81]

Defining Digital Preservation. A working group of the Preservation and Reformatting Section of the Association for Library Collections and Technical Services of the American Library Association (ALA) is drafting a standard operational definition for *digital preservation* that would be used in policy statements and other documents.[82]

Although the aforementioned tools and standards are instrumental, digital preservation programs at many libraries and cultural institutions are still in pilot or test modes. In a 2007 review of digital preservation readiness studies, Liz Bishoff stressed the importance of expanding educational opportunities for staff involved in preservation and curation programs.[83] Such opportunities are critical to integrate the digital preservation tools and emerging standards into daily practice.

---

[78] ISO 14721:2003 OAIS : http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=24683&ICS1=49&ICS2=140&ICS3.

[79] Audit Checklist for Certifying Digital Repositories: http://www.rlg.org/en/page.php?Page_ID=20769.

[80] Core Requirements for Digital Archives: http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=92.

[81] UK Digital Curation Centre (DCC) and Digital Preservation Europe (DPE). Digital Repository Audit Method Based on Risk Assessment. March 2007. Available at http://www.repositoryaudit.eu/.

[82] Defining Digital Preservation: http://blogs.ala.org/digipres.php.

[83] Liz Bishoff. 2007. "Digital Preservation Assessment: Readying Cultural Heritage Institutions for Digital Preservation." Paper presented at DigCCurr 2007: An International Symposium in Digital Creation, April 18–20, 2007, Chapel Hill, N.C. Available at http://www.ils.unc.edu/digccurr2007/papers/bishoff_paper_8-3.pdf.

4. Implications of LSDIs for Book Collections

The implications of LSDIs go beyond daily routines and digital preservation responsibilities to other areas within a library's operation. This section highlights the potential impact of LSDIs on book collections to demonstrate the ripple effect of mass digitization efforts.

4.1 Pressure for Relieving Space

Many research libraries face serious space shortages. In response to changes in library use, they are reducing the amount of space devoted to storing print materials in order to expand the user study and research areas. Will LSDIs affect libraries' decisions about how to use their physical space and how best to deal with their book collections? For instance, will there be more pressure from university or library administrations to eliminate duplicate copies of books or to store them off site? Also, what will happen to print originals after they are reformatted? In the era of microfilming, originals were sometimes discarded after being filmed. Likewise, it may be tempting to use the acquisition of a digital surrogate as a justification to deaccession original print material. What would be the long-term implications of discarding print copies on the basis of the existence of digital versions that may be incomplete or below-standard image quality?

Much of world's scholarly literature is not yet available in digital format. Consequently, research libraries continue to invest substantial amounts of funds in acquiring, cataloging, and housing print collections. This not only requires space but also strains collection development.

The precedent set by journal literature is an interesting one to analyze in regard to its potential implications for other library materials. According to a 2006 OCLC study, ARL members are rapidly accepting electronic format as the dominant medium for journal collections.[84] From 2002 to 2006, subscriptions to journals in print format decreased by 32 percent, whereas journals obtained in electronic format increased by 34 percent. As libraries move into a predominantly electronic-subscription environment, concerns about ownership and perpetual access to journal literature are growing.

4.2 Impact on Traditional Preservation and Conservation Programs

Preservation departments within libraries are responsible for the preservation of collections. This includes broad risk assessment and policy formulation and activities ranging from disaster preparedness and environmental control to single-item conservation treatment. Digital reformatting and digital preservation are under the purview of some preservation departments. Institutions vary as to how much of their "traditional" preservation activities are funded by grants

[84] Chandra Prabha. 2007. "Shifting from Print to Electronic Journals in ARL University Libraries." *Serials Review* 33(1) [March]: 4-13.

from public and private agencies, but some institutions rely heavily on external funding. The vast collections of digital resources made available by the LSDIs raise questions about how institutions will set priorities and allocate funds to support traditional preservation activities. Four issues come to mind:

1. Google argues that because only limited information will be available through snippets provided for in-copyright materials, book sales and library circulation of digitized material are likely to increase. According to an *Atlantic Monthly* article, Google is less likely to destroy the book business than to "slingshot it" into the 21st century.[85] If this is true, library circulation probably will increase, especially through interlibrary loan. Early evidence indicates that there are likely to be increases in usage. During a discussion at the 2007 ALA Annual Conference, some Google library partners reported an increase in interlibrary loan requests generated by use of Google Book Search.[86] Although broadening the reach and use of collections is desirable, such an increase in borrowing would also expose more books to wear and tear as they move among libraries. (Interlibrary lending usually subjects books to more damage and risk of loss than an ordinary circulation does.) In turn, this will require strengthening the preservation and conservation programs that maintain the artifactual integrity of materials. According to the 2004–2005 ARL Preservation Statistics, total preservation expenditures continue to be stagnant (i.e., not to keep up with inflation), which raises concerns about the ability of libraries to expand and fund these operations.

2. A number of funding agencies make grants available for preservation surveys, conservation treatment, and reformatting. Some of these funders may question the value of maintaining and preserving book collections that are available in digital format. If the value of preserving such print publications is not articulated and justified, funders may shift their priorities. One justification for retaining print copies is that they can be considered backups or "leaf masters"[87] for the digital copies. That is, when a page has been scanned poorly or a page is missing from the digital version, the original print copy can be referenced to remedy or elucidate the concern. A trend that strengthens the feasibility of this backup role of print is the increasing use of what is commonly called "offsite storage." Such spaces might more appropriately be referred to

[85] Michael Hirschorn. 2007. "The Hapless Seed." *Atlantic Monthly* 299(5): 134-139.

[86] "The 'Google Five' Describe Progress, Challenges." 2007. *Library Journal Academic Newswire* (June). Available at http://www.libraryjournal.com/info/CA6456319.html.

[87] Gary Frost, university conservator at the University of Iowa Libraries, introduced the leaf master idea, which implies a continuing role for originals within access and delivery systems. He argues that screen presentation of print serves a utility function by enhancing access but does not preclude the need to keep print originals available for consultation. More information on leaf mastering can be found at http://futureofthebook.com/storiestoc/leaf.

as "collections preservation centers" since the environmental and security conditions in such facilities are much better suited to the longevity of paper than are conditions in most libraries, where patron comfort must be taken into account.

3. The process of pulling from shelves, shipping, and digitizing in LSDIs puts the original at risk. Although there are no indications that current bound-volume scanning technologies are inherently damaging, there has been no systematic study of the impact of digitization on the physical condition of book collections. At the aforementioned 2007 ALA Annual Conference, a panel composed of Google Book Search participating libraries noted that damaged books are one of the challenges they face. Most of the institutions responding to the LSDI preservation survey indicated, however, that they have experienced no or minimal damage. The condition of materials prior to digitization will influence the risk of damage. If there is damage during digitization, there may be disagreement as to whether it was caused by improper handling or by years of storage in suboptimal conditions.

4. Since digitization was introduced to the library community in the early 1990s, librarians have discussed the future of the book as artifact and its contributions to the intellectual value that are difficult to capture through digital reformatting, such as the historical context provided by binding, watermarks, and chemical composition of ink. In 2001, the Council on Library and Information Resources (CLIR) convened a task force to investigate the role of the artifact in libraries and archives.[88] The group concluded that preservation budgets often fail to meet the preservation needs of artifacts. Members projected that increasing attention to digital reformatting "has the potential to eclipse the preservation needs of artifacts and to preoccupy the attention of the research community." Their main recommendation was to establish regional repositories to house and properly treat low-use print matter. A related suggestion was to convene a national committee to investigate the establishment of archival repositories that would retain a "last, best copy" of U.S. imprints. Both topics have been discussed over the past several years, without significant progress. Recent activities surrounding the North Atlantic Storage Trust, however, do show promise.

4.3 Print-on-Demand Books

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—espe-

---

[88] *The Evidence in Hand: Report of the Task Force on the Artifact in Library Collections*. 2001. Washington, D.C.: Council on Library and Information Resources. Available at http://www.clir.org/PUBS/reports/pub103/contents.html.

cially for longer materials such as books.[89] Some LSDI libraries are exploring the possibility of offering print-on-demand (PoD) services (especially for public domain materials) in cases where the individual contract allows it.

Although PoD issues do not relate directly to the main topic of this paper, they offer a good example of how today's decisions will affect future library programs.[90] 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 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.

Two LSDI libraries, the University of Michigan and Cornell University, are already using the PoD service provided by BookSurge, a subsidiary of Amazon, to make digital content created through institutional efforts available for online ordering. In June 2007, BookSurge and Kirtas technologies announced a collaboration with Emory University, the University of Maine, the Toronto Public Library, and the Cincinnati Public Library to digitize rare and inaccessible books from their collections and to distribute them through BookSurge's PoD service.[91] According to a press release from Emory, the digitization and digital publishing model allows the library to retain control of the digitized versions of its collections.[92] This includes exposing the full-text content for indexing by various search engines, rather than just the partnering Web company.

5. Recommendations

A primary incentive for libraries' participation in LSDIs is to provide broader and easier access to books through the popular search engines by aggregating supply and demand and enabling keyword-level searches. Although the current course of action may not be fully satisfactory, participating libraries maintain that without support

---

[89] 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 the difficulties reported with reading large amounts of text on a computer screen. Michael Levine-Clark. 2006. "Electronic Book Usage: A Survey at the University of Denver." *Libraries and the Academy* 6(3): 285-299.

[90] The information about the print-on-demand privileges provided to LSDI libraries for the digital copies that will be provided to them is considered confidential and is often included in contracts under a nondisclosure clause.

[91] BookSurge, an Amazon Group, and Kirtas Collaborate to Preserve and Distribute Historic Archival Books. June 21, 2007. Press release. Available at http://biz.yahoo.com/prnews/070621/nyth056.html?.v=85.

[92] Emory Partnership Breaks New Ground in Print-On-Demand Books. June 6, 2007. Press release. Available at http://www.news.emory.edu/Releases/KirtasPartnership1181162558.html.

from commercial entities, they would not be able to embark on such ambitious projects. Is this assertion strong enough to mitigate some of the concerns identified in this report? Should we perceive these ventures primarily as access projects, rather than as reformatting initiatives that yield high-quality digital surrogates for the original? If so, how can we define a preservation strategy that is built on this recognition?

Even the brief assessment presented in this paper shows that such questions are complex, interdependent, and open for interpretation. Formulating a joint action plan by the cultural institutions is desirable and will help clarify commonly debated aspects of LSDIs. It will be important to bring Google and Microsoft, as well as other commercial leaders, into this conversation. Participating libraries should take advantage of the partners' meetings organized by Google and Microsoft to present and discuss the community's digital preservation concerns and plans. However, it is important to acknowledge that there are institutional differences in opinion, digital library infrastructures, funding models, and strategic goals. In this context, the following recommendations aim to facilitate a discussion of the matter at hand. The recommendations center around five themes that weave through the LSDI preservation mandate: digitization as a potential method to preserve books (5.1–5.3); enduring access (5.4–5.5); preservation management (5.6–5.7); digital preservation strategies (5.8–5.9); and research library strategies (5.10–5.13).

5.1 Reassess Digitization Requirements for Archival Images

The prevailing digitization standards and best practices were established 15 years ago. They were created during a time of early implementations and were based on modest collection sizes and often on bitonal scanning. We need new metrics that are based on current imaging technologies, quality assessment tools, archiving practices, and evolving user needs. It is time to create new digitization metrics that take into consideration the following characteristics of the current landscape:

- contemporary digitization technologies and image-processing tools[93]
- ingest and storage guidelines and experience built over the past several years
- new archival file formats, such as JPEG2000 and PDF/A
- evolving access formats (such as XML) that are essential to support sophisticated retrieval and use of content such as text mining

[93] Burns and Williams present 10 principal image-quality attributes and represent current imaging science knowledge distilled to a simple form. See Peter D. Burns and Don Williams. 2007. "Ten Tips for Maintaining Digital Image Quality." Pp. 16-22 in *Archiving 2007*. Final proceedings of conference held May 21-24, 2007, Arlington, Va. Springfield, Va.: The Society for Imaging Science and Technology.

- the impact of lossy compression techniques and image processing on future preservation actions[94]
- the correlation between image quality and OCR accuracy
- the role, potential, and value of preservation metadata (PREMIS) and technical (NISO/ANSI Z39.87) metadata in supporting preservation actions
- requisite descriptive and structural metadata for supporting discovery and retrieval of digital materials

There are anecdotal data about the quality of images provided to participating libraries. It may be useful to have a systematic image-quality study based on inspection of sample images and associated metadata to evaluate the suitability of digital objects for preservation purposes. Such an assessment should be undertaken with two key considerations in mind. The first consideration is how to judge image quality in such an analysis. Should we rely on existing best practices, or should the evaluation be based on newly defined parameters suggested in this recommendation? The second factor is understanding the role of institutional missions and resources in defining preservation quality.

5.2 Develop a Feasible Quality Control Program

We need to reassess the quality control policies, tools, and workflows that were created to support small-scale digitization projects and to acknowledge that it is neither practical nor feasible to apply existing QC protocols to LSDIs. Williams has noted that today's ISO protocols for assessing digitization device performance are based on sound science and are quite reliable into the foreseeable future.[95] However, these QC targets and software were not designed to work in the high-volume, high-demand workflows of LSDIs. It is time to devise new models with calculated risks. Here are some ideas to expand thinking about our options:

- QC programs are important in ensuring the technical quality of digital content created by LSDIs. However, it is important to emphasize the importance of creating good-quality images during the initial capture so that QC is an assurance process to catch infrequent problems rather than a front-line strategy. The library community should negotiate rigorous technical specifications with digitization partners to reduce the pressure on the QC stage in catching missing or unacceptable images. The possibility of problems caused by equipment failure, digitization operator exhaustion, or use of uncalibrated equipment should be anticipated

[94] The National Library of the Netherlands is investigating the advantages and disadvantages of using compression within a long-term preservation context. The goal is to compromise between the need for reducing storage costs and the requirements for digital preservation, suggesting a more realistic approach to the long-term storage challenge. See Judith Rog. 2007. "Compression and Digital Preservation: Do They Go Together?" Pp. 80-83 in *Archiving 2007*.

[95] Excerpt from e-mail communication between the author and Don Williams, technical imaging consultant, standards and image quality, June 2007.

at the point of digitization, and appropriate measures should be in place to prevent them. A well-negotiated, well-developed QC program at the digitization center should enable the library to streamline the QC program it has in place to ensure that the digitizing partner is adequately meeting the agreed-upon quality.

- Traditional QC programs have four key components: development of quality parameters and QC methodology, identification of problem images (or other deliverables such as OCR or metadata), correction of problems, and integration of these improved objects into the collection. It is usually easier to identify problems than to fix them and to integrate them into a digital collection. Even if funding is not available to correct unacceptable images or other digital products, it is worth recording this information to support future actions. Such data can also be used to document how digital surrogates differ from original print materials; in this context, they can be considered a component of provenance information. Quality control is not necessarily a fixed process. Image-enhancement techniques continue to evolve and can be applied as they develop.

- Even a modest QC process will reveal errors that can be traced to problems at the point of digitization, such as poor performance of a specific piece of equipment or improper settings for image-processing applications. For such an approach to be effective, libraries would have to review files soon after they are received and report findings to digitization partners. Only in this way could patterns be investigated and their causes identified.

- Should we consider changing our approach to quality control, that is, applying it "as needed," rather than "just in case"? Such a strategy will involve relying on an automated image-quality process coupled with methods to promote receiving feedback from online readers. For example, the "Provide Feedback" link at the bottom of the Google Books page includes a form on which to report image-quality problems such as readability, completeness of page, curved or distorted text, and accuracy of bibliographic information. This may not be a comfortable practice for libraries; however, it may be a viable option given limited funding and QA concerns. With the as-needed approach, retaining leaf masters becomes more important.

- What would be the value and financial implications of a system in which participating libraries are responsible for filling in missing pages or rescanning unacceptable images? Sometimes the most complicated process in making corrections is reintegrating the corrected pages with the original materials. In this case, what is the incentive for the digitizing partner to meet the agreed-upon quality guidelines?

5.3 Balance Preservation and Access Requirements

Because of stakeholders' multiple perspectives, it has always been difficult to agree on a single digitization method that suits all circumstances. The LSDI institutions are recognizing that it is not feasible to fully adopt existing preservation digitization practices because of the scale of their endeavors. They are seeking compromises through various methods—dropping or reducing QC programs, settling for resolutions lower than 600 dpi, or switching to different file formats. The Society of Imaging Science and Technology's Archiving 2007 Conference featured several presentations on the impact of image compression and digital preservation. Presenters acknowledged that LSDIs must implement space-efficient digitization strategies to minimize long-term storage costs and to increase transmission efficiency for delivery and transfer.[96]

Librarians recognize the value of high-quality images to support a range of future needs, including preservation. They also recognize that the scale of the LSDIs makes maintaining the current best practices for high-quality images problematic. It is time to try to reach agreement about what is "good-enough" quality in LSDIs and to clarify what future needs they are intended to address. In seeking compromise, the library community should continue to advocate for high standards. As Cliff Lynch suggested during the 2006 Symposium on Scholarship and Libraries in Transition, digitization can provide a form of insurance for preserving content, even though digital surrogates cannot replace physical originals.[97]

5.4 Enhance Access to Digitized Content

Research libraries are making significant investments in archiving LSDI-generated collections. Such investments will be more worthwhile if discovery, access, and delivery are given equal emphasis. Digital content that is not used is prone to neglect and oversight; reliable access mechanisms are essential to the ongoing usability of these online materials. It is also important to reach out to new users and to expand tools for discovering and using digital information.
When asked about their plans for the digitized content they receive, some LSDI libraries say they will experiment with enhanced access and discovery tools and text-mining techniques. Achieving this ambitious goal will be possible only if libraries pool their resources and build on each other's accomplishments. An example of such a joint program (although modest in scale) is DLF Aquifer, which promotes

---

[96] See Stephen Chapman, Laurent Duplouy, John Kunze, Stuart Blair, Stephen Abrams, Catherine Lupovici, Ann Jensen, Dan Johnston. 2007. "Page Image Compression for Mass Digitization." Pp. 37-42 in *Archiving 2007*. See also Rog 2007, op cit.

[97] U.S. National Commission on Libraries and Information Science. *Mass Digitization: Implications for Information Policy*. Report from Scholarship and Libraries in Transition: A Dialogue about the Impacts of Mass Digitization Projects. Symposium held March 10-11, 2006, at the University of Michigan, Ann Arbor. Available at http://www.nclis.gov/digitization/MassDigitizationSymposium-Report.pdf.

effective use of distributed digital library content for teaching, learning, and research in the area of American culture and life.[98]

Noting that LSDIs to date have focused on keyword search to enhance discovery, Don Waters cautions the library community to consider sophisticated search and discovery methods, including new analytical techniques for content analysis.[99] As libraries assess various LSDI image and metadata quality parameters, it is critical to involve scholars in the process to incorporate their evolving requirements for viewing and studying digital content. CLIR and Georgetown University are conducting a project to assess the utility to scholars of several large-scale digitization projects, including Google Book Search, Microsoft Live Search Books, Project Gutenberg, Perseus, and the American Council of Learned Societies' E-Book project. CLIR expects to report on the project in mid-2008.

Copyright information is a critical element in making both preservation and access decisions. The Spring 2007 DLF Forum featured a session on building communities and systems for sharing and searching information about copyrights and their holders, especially within the context of LSDIs. This is a potentially rewarding area of collaboration for libraries. OCLC is defining core requirements for a collaborative copyright decision support tool that might help eliminate some of the system-wide redundancies. They are exploring the possibility of leveraging WorldCat, as it represents the shared investment of many libraries in aggregating various metadata for their print collections. MARC records do not contain all the information necessary to make a copyright determination and will need to be supplemented.

5.5 Understand the Impact of Contractual Restriction on Preservation Responsibilities

Google and Microsoft restrict the sharing of full-text digitized content. Participating libraries can, at best, share copies of digitized materials only with academic institutions and only as long as they agree not to make the files available to other commercial Internet search services. Such restrictions, which aim to prevent hosting of the digitized books by other commercial search engines, are likely to impede some preservation strategies, such as redundancy arrangements. Having more than one search engine host the same content

---

[98] The Digital Library Federation's Distributed Open Digital Library initiative was launched in 2003 to pool existing digital library assets, resources, and services. In 2006, the initiative evolved into DLF Aquifer with a refined focus on promoting effective use of distributed digital library content for teaching, learning, and research in the area of American culture and life. The project specifically addresses the difficulty humanities and social science scholars face in finding and using digital materials located in a variety of environments with an array of interfaces and protocols. The project is funded by The Andrew W. Mellon Foundation. DLF Aquifer: http://www.diglib.org/aquifer/.

[99] Don Waters' comments are included in Richard K. Johnson, "Google's Broad Wake: Taking Responsibility for Shaping the Global Digital Library." *ARL: A Bimonthly Report* 250 (February 2007). Available at http://www.arl.org/bm~doc/arlbr250digprinciples.pdf.

is likely to increase the survival of digital materials. A group of legal scholars, including Jack Lerner and Jennifer Urban, is conducting research on legal restrictions imposed by LSDI contracts with a focus on the Google initiative.[100] The library community will benefit from forming a united front to address with commercial partners the limitations that they place on their copies of digital materials.

5.6 Lend Support for Shared Print-Storage Initiatives

With the increasing value placed on online access, research institutions will be under pressure to justify investments in maintaining their legacy print collections, some of which are low use and redundant. Consolidation of holdings in a shared-storage environment can offer significant space savings as well as improved control of ambient temperature and humidity. Agreements among geographically distributed print repositories can create additional economies of scale. There is a long history of working groups and programs exploring shared print-storage solutions. However, efforts are often curtailed or stalled because of the complex and political nature of governance issues, especially in regard to types of materials, duplication, ownership, and funding for sustainability.

OCLC Programs and Research has embarked on a series of studies and programs aimed at identifying the key incentives and obstacles to institutional collaboration in this area.[101] In 2006 RLG Programs began working with members of the North American Storage Trust to develop a policy framework that would enable participating libraries to assess local print collections in light of ongoing community investments in off-site storage.[102] In fall 2007, OCLC Programs and Research published a report that examines the state of the art in library off-site storage, identifying gaps in the current infrastructure and new opportunities for community and institutional action.[103] Several regional initiatives are also in place. National and regional

---

[100] Jack Lerner is a visiting clinical assistant professor at the Gould School of Law at the University of Southern California (USC) and acting director of the Intellectual Property and Technology Law Clinic at USC. Jennifer Urban is clinical associate professor of law and director of the Intellectual Property and Technology Law Clinic at Gould School of Law at USC.

[101] OCLC/RLG Shared Print Collections Program: http://www.oclc.org/programs/workagenda/collectivecoll/sharedprint/.

[102] The planning efforts involve identifying critical governance issues such as joint commitments to retain and provide continuing access to locally owned research collections. The goal is to establish a network of print repositories bound by explicit community agreements to long-term preservation and access. Institutions participating in the network would commit to disclosing retention and access policies for discrete collections (e.g., materials held in off-site storage facilities) and would gain priority access to the collectively maintained preservation collection. Locally redundant holdings might then be reduced or eliminated in light of aggregate holdings and shared preservation and access commitments. OCLC/RLG Shared Print Collections Program: North American Storage Trust: http://www.oclc.org/programs/workagenda/collectivecoll/sharedprint/nast.htm.

[103] Lizanne Payne. 2007. *Library Storage Facilities and the Future of Print Collections in North America*. Report commissioned by OCLC Programs and Research. Available at http://www.oclc.org/programs/publications/reports/2007-01.pdf.

shared-storage efforts demonstrating strong leadership need firm support from the library community.

5.7 Promote the Use of Registry of Digital Masters

Developed in 2001, DLF's *Registry of Digital Reproductions of Paper-based Monographs and Serials* aimed to provide functional specifications for a registry that records information about digital reproductions of monographs and serials. It evolved into the DLF/OCLC Registry of Digital Masters (RDM), a central place for libraries to search for digitally preserved materials.[104] The premise of the RDM was that a library, by registering digitized objects, indicated that the digital copy was created under established best practices for digitization and that the institution was commited to its digital preservation. However, books digitized as part of the LSDIs do not necessariy adhere to established best practices. As noted in Section 5.1, it is necessary to revisit what is considered "acceptable quality." In addition, there is need to further articulate how institutions define a "commitment to digital preservation."[105] The registry has the potential to reduce redundancies and to record an array of relevant information that will support the preservation of content as well as the planning of future digitization efforts.

There is some debate about how much effort should be spent on reducing redundancy. Some argue that duplicating efforts is more cost-efficient than trying to manage a coordinated selection process. Another argument in favor of redundancy is that if one digital copy becomes corrupted or inaccessible, another will be available. Having more than one digital copy also increases the chance that there will be a better copy among the duplicates. Although each of these claims has merit, there are also compelling reasons for sharing information and minimizing redundancies, not the least of which is to be able to attract funding by identifying unique content.

Given the concentration on speed and production efficiency for LSDIs and the volume of materials processed, it is not realistic to attempt to locate missing pages or replacements for torn ones. It is possible for a digital collections registry to maintain information about such incomplete files, so that they are earmarked for future action or at least documented in authenticity and provenance metadata. Ideally, such a registry could also track quality problems with images, metadata, and OCR files.

Rather than relying on LSDI libraries to register digitized content, it may be more effective for OCLC to work with Google,

---

[104] DLF/OCLC Registry of Digital Masters: http://www.oclc.org/digitalpreservation/why/digitalregistry/.

[105] According to an August 2007 e-mail message to the author from OCLC's Susan Westberg, "the guidelines follow the basics of DLF digitization best practices, but whatever the institution chooses, to follow DLF standards or their own, they need to include a statement about or access to their digitization standards in the bibliographic record, to let other institutions determine whether or not those standards are high enough and objects don't need to be digitized again."

Microsoft, OCA, and MBP to automatically ingest and record such information, though it would be best to supplement that information with pointers to the university's digital copies. OCLC is working with Google and Microsoft to synchronize WorldCat with digitization efforts.[106] OCLC eContent Synchronization is designed to automatically create a record in WorldCat representing the digital manifestation.

5.8 Outline a Large-Scale Digitization Initiative Archiving Action Agenda

The most newsworthy aspect of the CIC's June 2007 announcement that it would join the Google Book Search initiative[107] was the consortium's decision to create a shared repository to jointly archive and manage the domain content, including producing customized discovery portals to meet the needs of each institution's user community. Although this model may not suit all LSDI institutions, it presents an option for those with limited resources or preservation programs.

In their survey of e-journal archiving, Kenney et al. concluded that academic libraries have been slow to address the vulnerability of e-journal literature because of competing priorities in their organizations and because of a lack of experience in collective and shared digital archiving.[108] The report recommends that archiving programs that meet the standards and best practices be certified as trusted digital repositories (when certification is available) and that they provide compelling public evidence that they are equipped to manage collections. Similar principles will apply to a joint LSDI digital archiving program. The archiving program could be structured among libraries that are using a cooperative approach such as LOCKSS (Lots of Copies Keep Stuff Safe) or by a third-party archiving program similar to Portico.[109] Analogous to the involvement of publishers in the LOCKSS and Portico efforts, it will be useful for Google and Microsoft to be brought into dialogues to jointly address the long-term viability of digitized materials. The e-journal archiving report also

---

[106] Author's personal e-mail communication with Bill Carney, OCLC, August 2007.

[107] CIC/Google Book Search Project: Frequently Asked Questions: http://www.cic.uiuc.edu/programs/CenterForLibraryInitiatives/Archive/PressRelease/LibraryDigitization/FAQ6-5-07finalREV2.pdf.

[108] Anne R. Kenney, Richard Entlich, Peter B. Hirtle, Nancy Y. McGovern, and Ellie L. Buckley. 2006. *E-Journal Archiving Metes and Bounds: A Survey of the Landscape*. Washington, D.C.: Council on Library and Information Resources. Available at http://www.clir.org/PUBS/abstract/pub138abst.html.

[109] LOCKSS is open-source software that provides libraries with an efficient way to collect, store, preserve, and provide access to their own, local copy of authorized Web published content (http://www.lockss.org/lockss/Home). Examples of LOCKSS cooperative preservation projects (based on Private LOCKSS Networks) are available at http://www.lockss.org/lockss/Related_Projects. The mission of Portico is to preserve scholarly literature published in electronic form and to ensure that these materials remain accessible to future scholars, researchers, and students. It offers a service that provides a permanent archive of electronic scholarly journals (http://www.portico.org/).

urges consideration of whether there should be a certification process to assess the ability and readiness of commercial partners to digitize the library collections.

Developing a common archival strategy is a complex process. Agreeing on key principles and endorsing a joint plan continues to be a stumbling block. A wide range of archival models and policies have been customized to individual institutions' goals, resources, and content types; furthermore, diversity of preservation strategies allows the library community to experiment and select the best of the approaches. Possibilities for collaboration extend well beyond providing a common preservation repository. Effective collaboration might also include the following:

- defining minimum digital preservation requirements necessary to ensure the persistence of digital materials and associated metadata files to facilitate shared storage and registry initiatives
- working with IT groups within cultural institutions (such as theory centers, central IT units, academic technologies, computer science departments) to develop and manage shared large-scale storage systems
- making data-redundancy arrangements among libraries for back-up, or implementing other distributed and collaborative strategies such as LOCKSS
- developing storage metrics to share configuration and cost information in standardized ways
- supporting standards for storage-management interoperability
- sharing open-source preservation applications and collaborating to develop access and preservation services as flexible and scalable components to be added to repository models supporting preservation activities
- exploring usage trends created by the online availability of materials to assess how the 80/20 rule applies in the digital world and to consider how usage statistics can inform preservation decisions in support of priority setting and risk taking[110]
- exploring how to incorporate risk assessment strategies in making and implementing preservation decisions, being sure to consider how preserving the analog books might affect the risk assessment strategies for the digital versions, and vice versa.[111]
- creating a wiki (or a similar collaboration tool) to systematically

[110] Cornell's sampling of usage data for some of its digital collections showed that about 40 percent of the downloads are drawn from 20 percent of the collection. The initial sale statistics for Cornell digital books offered through Amazon's print-on-demand option indicate that 13 percent of the titles have been ordered once or more. This is an early finding of sales with minimal marketing. However, every day there are new "first-time sales" of titles, indicating that aggregating supply creates demand even for unused library materials.

[111] Risk assessment and management within the context of digital curation and preservation is described in the Digital Repository Audit Method Based on Risk Assessment toolkit, available at http://www.repositoryaudit.eu/.

distribute up-to-date information about preservation strategies implemented by different libraries
- offering consultancies, workshops, and training sessions

5.9 Devise Policies for Designating Digital Preservation Levels

One of the findings of the Getty Research Institute survey was that, organizationally and financially, we cannot keep all digital content and preserve it at the same level of service and functionality. LSDI libraries must therefore determine the extent and type of their preservation efforts. Given that the library community is unlikely to have funds to redigitize the same content, digital books will inevitably be viewed as "insurance copies"—as backups for originals (regardless of the questions about quality). Because selection can be time-consuming and expensive; it is likely that the trend will be to preserve everything for "just-in-case" use.

There are two options with respect to preservation: (1) all files can be automatically preserved at the same level; or (2) metrics may be used to make a decision on the basis of the material's perceived value and use. For example, storage-redundancy arrangements may be implemented only for content that is considered of high scholarly value. This topic is well worth exploring further by means of a risk analysis of cost-efficient preservation strategies for low-use content.

Finally, as cultural institutions explore assessing preservation levels based on perceived scholarly value, it is important to consider the implications of such decisions on the breakthroughs that result as scholars rediscover and repurpose information that has been long out of use. Judging the scholarly value of library materials is a complicated and subjective process; nevertheless, it will be a stimulating undertaking for the library community.

5.10 Capture and Share Cost Information

Some LSDI libraries indicate in their FAQs and press releases that their commercial partners are digitizing content at their own expense.[112] It is true that digitization costs such as materials shipping, scanning, processing, OCR creation, and indexing are covered by commercial partners. However, staff members at participating libraries are supporting these initiatives by spending 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.

[112] The costs are across the board as both Google and Microsoft are including reimbursement provisions under nondisclosure agreements. According to the University of Michigan Library/Google Digitization Partnership FAQ, all costs related to pulling and reshelving materials are borne by Google. Often, information is concealed by the use of language; for example, Harvard's FAQ states that "Google is bearing the direct costs of digitization" (see http://hul.harvard.edu/hgproject/faq.html).

Cornell University Library currently invests close to seven full-time equivalent staff (distributed among a total of 25 staff members) in managing LSDI-related tasks for digitizing 10,000 books a month. It is difficult to calculate a fixed cost because of individual factors that affect selection and material-preparation workflows and the varied physical environments at participating institutions. Different staffing configurations are also required for ramp-up versus ongoing processes. It is important to document and acknowledge *all* the expenses for *all* the partners associated with LSDIs. Often neglected or underestimated in cost analysis are the accumulated investments that libraries have made in selecting, purchasing, housing, and preserving their collections.

It is difficult to identify the proportion of participating library contributions in overall LSDI expenses. Moreover, because of varying estimates of digitization costs, it is impossible to forecast the digitization investments of participating commercial partners. A quick review of the literature reveals no consensus on metrics or factors for calculating all the costs involved in digitizing a book. For example, the CIC's Google Initiative FAQ estimates the costs of digitization for the libraries before joining the Google program at about $100 per volume.[113] The Internet Archive claims that its digitization process costs about 10 cents a page, or $30 for a 300-page book.[114] These are not inclusive totals and may not include several pre- or post-processes.

One characteristic of LSDI agreements is that participating institutions maintain full rights over print materials that have been digitized. Faced with criticism about poor image quality, some institutions have suggested the possibility of rescanning the same content in the future—perhaps using institutional funds that would give them the freedom to set their own quality parameters. Is that a cost-effective, realistic alternative?

5.11 Revisit Library Priorities and Strategies

LSDIs have been unexpected and disruptive—at least for some of the participating libraries. The initiatives began at a time when research libraries were exploring their futures in light of developments such as Google's search engine for information discovery and a growing focus on cyberinfrastructure and the structures that support data-intensive initiatives. Libraries have been increasingly pressured to focus digital preservation efforts on the unpublished and born-digital information domain, where preservation concerns are most urgent. It will be tricky to balance the need to preserve the digital versions of already-published analog materials with the growing need to focus on born-digital materials.

---

113 CIC/Google Book Search Project: Frequently Asked Questions: http://www.cic.uiuc.edu/programs/CenterForLibraryInitiatives/Archive/PressRelease/LibraryDigitization/FAQ6-5-07finalREV2.pdf

114 Barbara Quint. 2005. "Open Content Alliance Expands Rapidly; Reveals Operational Details." *Information Today*, October 31. Available at http://newsbreaks.infotoday.com/nbreader.asp?ArticleID=16091.

Likewise, it will not be easy to deal with multiple and sometimes competing priorities in regard to access-related library projects. Although research and practice show that users increasingly prefer digital information and services, academic and research libraries remain under pressure to continue traditional services.[115] For example, recent user studies at Cornell indicate a growing demand for extended library hours. It is rare to hear about a service being eliminated in order to shift funds into a newly growing area. But the costs of processing and archiving new digital material may cause a significant shift in how funds are distributed among services at many libraries. It is important to try to define the LSDIs' relative role within the broader scope of library activities and mid-term strategies.

5.12 Shift to an Agile and Open Planning Model

One virtue of LSDIs is that the contributing libraries are gaining experience in interacting and negotiating with commercial information organizations, which function very differently than do academic institutions. As John Voloudakis has noted, today's need for faster responsiveness has introduced the "adaptive organization" strategic planning model for IT.[116] This model is characterized by an institutional focus on sensing and responding to the evolving environment as quickly as possible. In today's fluid IT environment, traditional strategic planning and consensus models are unlikely to support the decision-making processes of research libraries. There will be increasing pressure for quick responses to opportunities and changes. It will also be essential that libraries develop scalable and flexible infrastructures that facilitate rapid execution. Equally important is learning to take calculated risks. The summary of discussions at Digital Preservation in State Government: Best Practices Exchange 2006 notes that there are no "best practices" for digital preservation.[117] Instead, there are merely "good-enough" solutions. Holding out for an ideal solution is often not feasible; moreover, implementing less-than-perfect solutions can enable institutions to be flexible, modular, and nimble so that they can continue to refine their strategies as new options become available.

---

[115] The New York University Library's study of faculty and graduate student needs for research and teaching conclude that across disciplines there are widely differing expectations of the roles of the library. Many scholars still care deeply about the traditional roles of the library. Cecily Marcus, Lucinda Covert-Vail, and Carol A. Mandel. 2007. *NYU 21st-Century Library Project: Designing a Research. Library of the Future for New York University*. Available at http://library.nyu.edu/about/KPLReport.pdf.

[116] John Voloudakis. 2005. "Hitting a Moving Target: IT Strategy in a Real-Time World." *Educause Review* 40(2) [March/April]. Available at http://www.educause.edu/apps/er/.

[117] Christy E. Allen. 2006. "Foundations for a Successful Digital Preservation Program: Discussions from Digital Preservation in State Government: Best Practices Exchange 2006." *RLG DigiNews* 10(3) [June 15]. Available at http://www.rlg.org/en/page.php?Page_ID=20952.

5.13 Re-envision Collection Development for Research Libraries

In October 2005, Cornell University Library hosted the Janus Conference, a meeting at which participants explored the role and future of collection development in research libraries.[118] Several recommendations came out of the conference. For example, conferees called for coordination of a mass digitization project to facilitate retrospective conversion of research library collections (complementing the Google initiative) and urged research libraries to put in place a network of digital repositories that operate according to certified standards. Conferees also endorsed the creation of regional print repositories.

Although the forum was instrumental in pooling ideas and energy, there has not been much progress in advancing the agenda. The issues raised during the Janus Conference continue to be critical in determining the future of research library collections.[119] At the heart of many LSDI-related questions is the future direction for collection development programs in research libraries, and especially, how future selection and acquisition decisions will be shaped in the light of increased online content and worldwide access to core collections.

6. Conclusion: Why Join Forces?

Many of the recommendations set forth in this paper will require collaboration among cultural institutions. Whether or not current conversion efforts fully adhere to digital reformatting requirements, they are enormously resource intensive, and the library community needs to develop a plan to leverage the outcomes of ongoing digitization efforts. Meanwhile, as new partnerships are formed and technical and procedural guidelines for existing collaborative efforts are revised, it is important to continue to negotiate to raise the image- and metadata-quality bars.

Teamwork is a prevalent concept in the library community, but experience has shown that effective collaboration is hard to achieve. Ross Atkinson has pointed out that this is partially because the system usually "works" somewhat effectively, regardless of the success of collaborative efforts, and "writing and speaking about cooperation are viewed as forms of leadership, while the act of cooperating is not."[120]

One of the key requisites for collaboration is identifying a leader to coordinate agenda setting and implementation in addition to overseeing the assessment of outcomes. Currently, there is not a single US agency or an institution with the mandate of providing coordina-

---

[118] Janus Conference on Research Library Collections: Managing the Shifting Ground Between Writers and Readers. October 2005. Available at http://www.library.cornell.edu/janusconference/januskeys.html.

[119] Ross Atkinson. 2005. Introduction to the Break-Out Sessions: Six Key Challenges for the Future of Collection Development. Remarks delivered at the Janus Conference, Cornell University, Ithaca, N.Y., October 2005. Available at http://dspace.library.cornell.edu/bitstream/1813/2608/1/Atkinson_Talk.pdf.

[120] Ibid.

tion and leadership in the digital preservation domain. In the United Kingdom, JISC develops partnerships to enable UK education and research communities to engage in national and global collaborations to overcome the challenges of delivering world-class information and communication technology solutions and services. In the United States, several cultural and educational organizations and private foundations try to encourage partnerships through their initiatives and funding programs. However, there continues to be need for one or more institutions to assume the leadership role to facilitate alliance building among cultural institutions. The success of any collaborative effort requires the involvement of all stakeholders.

While cooperative initiatives have not come easily to libraries, there are some successful examples. For instance, preservation microfilming was a rewarding collaborative effort for several decades. One of the operating principles of this venture was adhering to uniform guidelines, and the effort was built on trust and mutual interest. Libraries will join forces in pursuit of a common agenda only when the benefits of collaboration outweigh the costs and when they see collaboration as a win-win situation. Reading some of the recommendations in this paper, one may rightfully ask, "What makes the LSDI agenda appealing enough to overcome the barriers to collaboration, and what are the incentives to work together?"

Stewardship Responsibilities. Cultural institutions have an obligation to protect the future of our scholarly heritage as a public good. Some library staff and scholars ask whether we should entrust our cultural heritage to partners with commercial interests simply for the sake of speed and expediency. This is a valid question. The library community needs to demonstrate its ability to fulfill its stewardship role, which should not preclude taking advantage of financial opportunities offered by commercial partners.

Enduring Access. The 800-pound gorilla in the LSDI preservation agenda is the future of Web access to digitized books. Many worry that digital content may no longer be available in the future through present-day search engine portals, which evolve rapidly in terms of both content and retrieval technologies. In a 2004 CLIR publication, Abby Smith stated that, "the fundamental purpose of preservation will be to ensure access to information to some user at some point in the future."[121] LSDI libraries may be in a position to take care of bit preservation at an institutional level and to use the digital copies for backups. However, providing enduring access by enabling online discovery and retrieval of materials (within limitations of copyright laws) for future generations is an enormous challenge—one that may not be met unless faced collectively. Efforts at the individual library level will not adequately address the enduring-access challenge unless there is a plan for providing aggregated or federated access to digital content. Today's users prefer searching

[121] Abby Smith. 2004. Mapping the Preservation Landscape. Pp. 9-16 in *Access in the Future Tense*. Washington, D.C.: Council on Library and Information Resources. Available at http://www.clir.org/PUBS/reports/pub126/contents.html.

and retrieving information in integrated search frameworks; they use digitized books only if they can be conveniently accessed in their preferred search environments and support their searching and reading preferences. Therefore, hosting public domain digitized books solely through individual library portals is likely to be insufficient.

Cost-Effectiveness. Although most digitization costs are borne by the commercial partners, the participating libraries are contributing substantial effort in preparing and managing content. The value of years of investment in purchasing and managing book collections is often underestimated. The LSDI flurry caught the library community at a time when many institutions were beginning to plan or develop digital preservation programs. Although the library community has some familiarity with digital preservation strategies, the quantity of data output from LSDIs dwarfs experiences to date. Not all libraries have the resources to assume a long-term archiving role for such large quantities of content. Shared-storage management is an example of such a cost-effective strategy. Cooperative arrangements can be achieved at many levels—through collaborations among libraries, through individual libraries working at their own home institutions with other related service providers, or both.

Future of Research Libraries. It is critical that the research libraries assess incentives for and impediments to collaboration from a broader perspective by taking into consideration emerging trends in research libraries. Libraries clearly need to modify their roles and programs to meet the needs of 21st-century users. The symbolic role of the library as the "heart of the university" is being challenged, and it is likely that different measures will be used to assess the role of libraries within an academic community.[122] The ARL is exploring how to modify the current practice of assessing research libraries on the basis of traditional quantitative measures such as collection size. One of the indicators of success among cultural institutions should be their willingness to contribute to joint agendas.

Joining forces among cultural institutions—ideally including corporate partners and content creators—will leverage resources, strengthen causes, control risks, and expand alternative strategies. Admittedly, there are institutional differences in opinion, funding models, digital library infrastructures, and strategic goals; consequently, not every action agenda lends itself for fruitful partnership. It is essential that cooperative efforts do not slow the community's efforts, but rather complement ongoing institutional programs.

It is time to have an open dialogue on a collaborative preservation agenda to determine which domains require independent or complementary programs with robust communication and to explore which tasks lend themselves to collaboration in the best interests of participating libraries, the library community in general, and current and future users.

---

[122] For example, see the findings of the following investigation on how the attitudes of university presidents and provosts towards their academic libraries have changed: Beverly P. Lynch, Catherine Murray-Rust, Susan E. Parker, et al. 2007. "Attitudes of Presidents and Provosts on the University Library." *College & Research Libraries* 68(3) [May]: 213-227.

APPENDIX

Large-Scale Digitization Initiatives: Survey of Preservation Implications

In July 2007, a Web-based survey questionnaire was distributed to 20 research libraries in the United States, the United Kingdom, and Canada. The goal of the survey was to gather information about the preservation activities of large-scale digital initiatives (LSDIs). The survey was distributed only to libraries that were actively participating in the Google, Microsoft, or Open Content Alliance (OCA) initiatives as of July, and was not sent to those who had signed agreements but were still in the planning stages.[123] To maintain the anonymity of respondents, we do not identify which libraries completed the survey; however, all are among those listed on page 7.

Fourteen of the 20 institutions were able to provide information about their large-scale digitization efforts. Six libraries were not able to participate for a variety of reasons, including privacy concerns and insufficient experience.

1. LSDI Participation

The two tables below summarize the distribution of respondents' participation in LSDIs. As the tables show, many respondents participated in more than one initiative.

Combined totals	
Answer Options	**Response Count**
Google	11
Microsoft	3
OCA	5

Breakdown of above totals	
Answer Options	**Response Count**
Google only	6
OCA only	1
Microsoft and Google	2
Google and OCA	3
Microsoft and OCA	1
Google, OCA, and Microsoft	1

[123] Because of the geographically distributed nature of the Million Book Project, the survey did not include the MBP participants. Several MBP project partners contribute only to the digital library research and development agenda. The MBP-related information presented in Section 2.3.2 of the paper was provided by the Carnegie Mellon University Libraries.

The total number of materials digitized by eight of the fourteen participating libraries was 22 million. The other six libraries did not respond to this question in quantitative terms; they characterized their selection efforts as evolving and were not able to quantify the number of materials digitized or slated for digitization. For example, one respondent commented that because of the nature of its catalog records and the period of material being considered for digitization (nineteenth-century), it was difficult to determine in advance the number of items that would fall within the scope of the project.

Seven libraries included both in-copyright and public domain materials in their LSDIs; the other seven included only public domain content.

When asked about the duration of the project, nine institutions' responses fell within the one- to six-year range. The others replied that the duration of their projects was undetermined or that the response to the question was confidential.

2. Digital Preservation Plans

Thirteen of the fourteen respondents expressed their intent to archive their digitized materials, that is, to assume long-term responsibility for preserving their digitized books. Twelve libraries said that their efforts were in the exploratory or planning stages; the other two libraries characterized their preservation efforts as "plans in place." Nine libraries indicated that they are developing a plan to ingest, store, and archive digitized content. Three libraries identified their repositories as ready to ingest, store, and archive.

When asked about collaboration in preservation efforts, two institutions stated that they already have partnerships in place, five institutions do not have any immediate collaboration plans, and four institutions indicated that they were considering a collaborative approach. Three institutions did not provide information in response to this question.

3. Challenges Ahead

Thirteen respondents commented on the challenges they faced. Many emphasized that the scale and pace of their LSDI require extremely robust systems, effective and reliable tracking tools, and tested preservation ingest procedures. Seven respondents stressed the difficulty associated with storing large amounts of data. The following comments illustrate the challenges perceived by the respondents:

- One library plans to base its preservation infrastructure on FEDORA architecture, but it has not yet tested it with such a large quantity of individual files.
- In regard to storage, one respondent observed that "the obvious challenges have also been the most basic." This library had found it very difficult to determine storage needs in advance.
- Creating and storing 4–8 gigabytes of data daily has put an enormous stress on one library's networking and storage system.

Several respondents stressed the time-consuming nature of data transfer.

- Three respondents expressed concerns about the lack of a clear institutional plan covering how long and why the library would be archiving the digitized books.
- The biggest challenge, according to one respondent, is the unproven state of preservation standards. This institution would like to be certified as a trusted digital repository, but at this point it does not perceive it possible "because the criteria are not realistic (as acknowledged by the group that developed and just revised them!)."
- Three libraries cited mischaracterization of mass digitization as preservation reformatting as a key challenge. They emphasized that the LSDIs were aiming at access, not preservation. One respondent noted that the resulting digital content may meet some preservation needs as well.
- One participant expressed concern about the quality of some items reformatted through mass digitization programs and noted that some of the digital content was not suitable to be used with evolving viewing technologies.
- Two respondents mentioned the impact of LSDIs on traditional preservation and conservation efforts. They indicated that an LSDI may draw attention to preservation needs that were not being addressed through mass digitization.
- Several respondents expressed concern about long-term financial challenges and the cost of the archival efforts.
- Appraisal and selection issues and the cost-effectiveness of maintaining duplicate copies of digitized content, especially given the current financial climate and competing priorities, were additional topics of concern.

4. Technical Requirements for Digitization

When asked to share imaging specifications (e.g., resolution, bit depth, file format, use of image-quality targets) for the digital copies they will archive, six libraries declined to provide information because of confidentiality obligations. Several libraries participating in the Google Initiative said that they have the "same specifications as for all other Google partners." Among the eight libraries that were able to provide information, one described its requirement as 600 dpi, 1-bit TIFF; the rest characterized their technical parameters as 300–400 dpi, 8–12 bit JPEG or JPEG2000.

The respondents were also asked to provide information about metadata standards used for description, structuring, and preservation. Of 10 libraries providing information, all listed MARC or MARC XML as their primary standard for descriptive metadata. Seven of these libraries are also using METS and are considering including MODS descriptive records. Three libraries are capturing MIX using JHOVE.[124]

[124] Information about the metadata standards referenced in this section is available at Standards at the Library of Congress: http://www.loc.gov/standards/.

5. Quality Control

One section of the questionnaire concerned inspecting the quality of digitized images received from a vendor. Eleven libraries indicated that they have a quality control (QC) strategy in place for this purpose. However, with one exception, they characterized their QC programs as evolving and noted the challenges faced because of the ambitious scale of digitization and limited resources. Their comments revealed a wide range of QC implementations, depending on institutional resources and initiative parameters. For example, one respondent said that his library inspects approximately five percent of newly digitized books for image quality and checks all files to ensure that they open. Checksums run as files are transferred to other media. Most of the responding libraries qualified their QC efforts as "small-sample based" and referred to their QC processes as "spot checks." Three institutions did not provide information about their quality control programs because of nondisclosure agreements.

When asked about the procedures for images or other associated deliverables, such as optical character recognition (OCR) files, with unacceptable quality, six (all Microsoft and/or Open Content Alliance participants) respondents indicated that the digital objects were sent back to the digitization service provider for correction. Three libraries recorded problems (two shared this information with the imaging center) but did not ask the service provider to make corrections. Five respondents said that they were either in the process of making decisions on this issue or that they could not share the information because of confidentially obligations.

6. Condition of Materials

One survey question aimed to elicit respondents' experience with respect to the physical condition of materials during digitization. Nine institutions checked "no or minimal damage," four had no opinion, and one respondent expressed concern about the level of damage. Some institutions that reported minimal harm noted that damage was not more than that experienced through normal use. One library conducted a pre- and post-condition survey early on and found no damage or minimal damage to its materials. The library that expressed concern stated that some books would be prone to damage regardless of how carefully they were handled. In keeping with curators' or preservation librarians' decisions, some libraries disbind books in which the text runs into the gutter and books that cannot be opened 180 degrees.

7. Completeness of Digitization Process

Asked whether they were tracking information about the completeness of the digitization process (e.g., missing pages, undigitized foldouts), six libraries replied that they were considering recording such information. Five libraries already had a system in place to capture such information, and one of them described an ongoing inventory

database development effort to record why books were rejected for scanning. This database will also support collection development efforts. Another respondent recognized that a certain percentage of books with errors and missing pages will be discovered only upon access by users and questioned how corrections will be managed for requests received from users. Three libraries were not able to share information because of nondisclosure terms.