DOCUMENT RESUME

ED 420 323 IR 057 122

AUTHOR Hazen, Dan; Horrell, Jeffrey; Merrill-Oldham, Jan
TITLE Selecting Research Collections for Digitization.
INSTITUTION Council on Library and Information Resources, Washington, DC.
ISBN ISBN-1-887344-60-2
PUB DATE 1998-08-00
NOTE 29p.
AVAILABLE FROM Council on Library and Information Resources, 1755 Massachusetts Ave., N.W., Suite 500, Washington, D.C. 20036 ($15).
PUB TYPE Guides - Non-Classroom (055) -- Reports - Descriptive (141)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Copyrights; Costs; *Decision Making; Information Sources; *Library Administration; *Library Automation; *Library Collection Development; Library Planning; *Research Libraries; Strategic Planning; Users (Information)
IDENTIFIERS *Digital Technology

ABSTRACT

This document proposes a model of the decision making process required of research libraries when they embark on digital conversion projects. A series of questions are offered that focus on facilitating the decision making process for library managers. Questions of what and how to digitize are placed in the larger framework of collection building by focusing, first on the nature of the collections and their use, and second, on the realities of the institutional context in which these decisions are made. This booklet is divided into 10 main sections: (1) Introduction; (2) Copyright: The Place To Begin; (3) The Intellectual Nature of the Source Materials; (4) Current and Potential Users; (5) Actual and Anticipated Nature of Use; (6) The Format and Nature of the Digital Product; (7) Describing, Delivering, and Retaining the Digital Product; (8) Relationships to Other Digital Efforts; (9) Costs and Benefits; and (10) Conclusion. "Selection for Digitizing: A Decision-Making Matrix" is also included on the back cover. (AEF)

Selecting Research Collections for Digitization

by Dan Hazen, Jeffrey Horrell, Jan Merrill-Oldham

August 1998

Council on Library and Information Resources

Commission on Preservation and Access
Digital Libraries
Economics of Information
Leadership

Selecting Research Collections for Digitization

by Dan Hazen, Jeffrey Horrell,
Jan Merrill-Oldham

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

August 1998

About the Authors

Dan Hazen, Librarian for Latin America, Spain, and Portugal at the Harvard College Library, is an area specialist actively involved in preservation microfilming and digital conversion projects both in America and abroad. Jeffrey Horrell, Associate Librarian of Harvard College for Collections, has served in library positions at the University of Michigan, Dartmouth College, and Syracuse University; he has also pursued research in the history of photography. Jan Merrill-Oldham, Malloy-Rabinowitz Preservation Librarian, directs the work of the Harvard University Library Preservation Center and the Harvard College Library Preservation Services Department, where appropriate means of integrating digital capabilities with established preservation and access strategies are being sought.

Commission on Preservation and Access

The Commission on Preservation and Access, a program of the Council on Library and Information Resources, supports the efforts of libraries and archives to save endangered portions of their paper-based collections and to meet the new preservation challenges of the digital environment. Working with institutions around the world, the Commission disseminates knowledge of best preservation practices and promotes a coordinated approach to preservation activity.

Digital Libraries

The Digital Libraries program of the Council on Library and Information Resources is committed to helping libraries of all types and sizes understand the far-reaching implications of digitization. To that end, CLIR supports projects and publications whose purpose is to build confidence in, and increase understanding of, the digital component that libraries are now adding to their traditional print holdings.

ISBN 1-887334-60-2

Published by:

Council on Library and Information Resources
1755 Massachusetts Avenue, N.W., Suite 500
Washington, D.C. 20036

Additional copies are available for $15.00 from the above address. Orders must be prepaid, with checks made payable to the Council on Library and Information Resources.

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.

Contents

FOREWORD

Collection building in the digital era presents challenges that libraries and archives have never before faced. They vary from having to work within licensing agreements in order to acquire serial publications, to having new, not yet well-defined options for providing service of analog items through digital conversion and dissemination. What role does the digitization of research collections play in a library's efforts to provide resources to its patrons when, where, and how they prefer to use them?

This paper proposes a model of the decision-making process required of research libraries when they embark on digital conversion projects. It is one of a series by CLIR dedicated to selection policy questions that have arisen in the digital information environment. The authors of the paper offer a series of questions to be answered that will facilitate the decision-making process for library managers. They place the questions of what and how to digitize into the larger framework of collection building by focusing, first, on the nature of the collections and their use, and, second, on the realities of the institutional context in which these decisions are made. Their method is, above all, most helpful in its pragmatic approach to the unsettling dynamism of the digital technology itself. They view technology as a tool to serve specific collections-related goals and assess the available technology for its ability to aid or obstruct access and preservation.

SUMMARY

Selection for digitization is a complicated process having much in common with selection for purchase, microfilming, and withdrawal, and with other strategic decision-making that is integral to the work of libraries. The conversion of textual, visual, and numeric information to electronic form—from preparation and conversion to presentation and archiving—encompasses a range of procedures and technologies with widely varying implications and costs. The judgments we must make in defining digital projects involve the following factors: the intellectual and physical nature of the source materials; the number and location of current and potential users; the current and potential nature of use; the format and nature of the proposed digital product and how it will be described, delivered, and archived; how the proposed product relates to other digitization efforts; and projections of costs in relation to benefits.

Copyright assessments play a defining role in digitization projects and must be addressed early in the selection process. If a proposed digitizing project involves materials that are not in the public domain, permissions must be secured and appropriate fees paid. If permissions are not forthcoming, the materials cannot be reproduced and the focus of the project must change. We will be able to convert to electronic form only a small percentage of existing scholarly materials, and to do even that will require substantial investments. Therefore, the intellectual value of the original sources, together with the types and levels of use, must shape priorities for conversion. Ideally, the electronic version of a source will permit new kinds of use and more sophisticated types of analysis. Decisions to digitize must also take into account the physical size, nature, and condition of source materials as they affect the characteristics of the desired product. Decisions must be based on the current state of technology, but they must also anticipate how changes in technology could enhance or make obsolete an investment in digitization. One must also assess how the product will be described for users, delivered to them, and managed over time.

Digitization, like other reformatting endeavors, takes place within a context larger than a single institution, discipline, or country. Selection decisions should be informed by both duplicative and complementary efforts. This may prove challenging, because it is difficult to determine whether an item has been already digitized and by what means. Cost-benefit analysis for digital conversion may also be hard to conduct reliably, because the costs of creating electronic resources vary considerably. File size, associated storage needs, and processing requirements account for part of the differences, though labor requirements are even more important. Functions such as

preparation of materials for scanning, indexing, bibliographic description, post-scan processing, and long-term file management often fail to be factored into cost equations. Incomplete cost analyses can impute benefits that are difficult to represent on a project balance sheet. Though digitizing projects must calculate the likely costs and benefits, our ability to predict either of them is as yet rudimentary. Thus, the decision to digitize must begin with an inquiry into copyright and an assessment of the nature and importance of the original source materials, but it must then proceed to analyze the nature and quality of the digitizing process itself—how well relevant information is captured from the original, and then how the digital data are organized, indexed, delivered to users, and maintained over time.

Authors' Acknowledgments

This essay grew out of the work of a Harvard University Library task force appointed late in 1995 and charged with drafting a broadbased white paper to help Harvard's librarians and curators plan digital projects. Rapid developments in digitizing and processing technologies, file naming and metadata creation, interpretation of copyrights and management of permissions, archiving, and other technical and administrative issues resulted in the emergence of selection as the topic that could be most effectively addressed at this time.

The essay owes much to Barbara Graham, Associate Director for Administration and Programs in the Harvard University Library, who convened the task force. Special thanks are due to colleagues Stephen Chapman, Preservation Librarian for Digital Initiatives, Lee Anne George, formerly Librarian for Information and Document Delivery Services in the Harvard College Library and now Program Planning Officer at the Association for Research Libraries, and Robin McElheny, Associate Archivist for Programs in the Harvard University Archives, who were members of the task force and made helpful comments on successive drafts. Stephen also collaborated on development of the accompanying flow chart.

INTRODUCTION

Electronic resources are immensely appealing to nearly everyone concerned with education and scholarship. The potential benefits of information in digital form—unfettered access, flexibility, enhanced capabilities for analysis and manipulation—are profound. The widely held notion that existing collections of books, manuscripts, photographs, and other materials should (and will) be digitized wholesale is not surprising. In reality, of course, the creation and maintenance of electronic resources require funding, skill, and ongoing commitment. Those that are intended for permanent use, moreover, will almost certainly require repeated intervention to ensure that they remain viable as technologies evolve. In creating digital products, libraries are called upon to balance the competing worlds of boundless promise and limited resources. Because hard choices are unavoidable, the decision-making process must be well organized and its results fully consonant with the institution's goals and values.

Selection for digitization is a complicated process having much in common with selection for purchase, microfilming, and withdrawal and with other strategic decision-making that is integral to the work of librarians and curators. Conversion of textual, visual, and numeric information to electronic form, however, involves additional layers of complexity. The digitization process, from preparation and conversion to presentation and archiving, encompasses a range of procedures and technologies with widely varying implications and costs. Digital reformatting of library collections is still in its infancy, at once limiting what can be accomplished now and forcing decision-makers to anticipate future improvements. Scanned images optimized for viewing on today's computer monitors, for example, will display poorly on tomorrow's high-resolution screens and will require reprocessing. The same may ultimately be true of bitmap texts, which, if they are not made word-searchable once conversion is affordable, may be underutilized by researchers who have come to rely on key-word search capability. Considerations such as these make selection for digitizing more challenging than selection for purchase.

The judgments we must make in defining digital projects require consideration of many factors, including: assessment of the intellectual and physical nature of the source materials; the number and location of current and potential users; the current and potential nature of use; the format and nature of the proposed digital product and how it will be described, delivered, and archived; how the proposed product relates to other digitization efforts; and projections of costs in relation to benefits.

COPYRIGHT: THE PLACE TO BEGIN

There are many interdependent and interacting factors to be weighed in selecting materials to digitize. The specific choices that result from the selection process will reflect subjective judgments, any of which may change over time. Nuanced assessments, ambiguity, and shades of gray are all to be expected.

Questions concerning copyright, however, are far more clear-cut. Simply stated, if a proposed digitizing project involves materials in the public domain, the work can proceed. If the source materials are protected by copyright but rights are held by the institution or appropriate permissions can be secured, the work can move ahead. If permissions are not forthcoming for copyrighted sources, however, the materials cannot be reproduced and the focus of the project must change. Copyright assessments thus play a defining role with regard to digitizing projects. Since the impact of copyright is so decisive, we have given it pride of place in this discussion.

Copyright issues in the digital environment are still very much in flux and have provoked ongoing international discussion. While the broad thrust of digital technology is toward enhanced access, diminished costs, and more versatile capabilities, it is far less clear that copyright law will likewise encourage wider use. The legal strictures applicable to a particular project will vary depending on the country in which the project is based, the country in which the source materials were produced, and prevailing international agreements. Different kinds of materials, moreover, usually pose different types of rights-management issues. The performance rights associated with musical scores, for example, or exhibition rights for films, differ from rights for nonperformance materials such as electronic journals or documentary photographs. To complicate matters, all these rights are susceptible to change over time.

Digital projects must be undertaken with a full understanding of ownership rights, difficult as they often are to ascertain, and with full recognition that permissions are essential to convert materials that are not in the public domain. Rights that must be negotiated with the copyright holder often entail fees. The institution hosting a project may also have policies and procedures that inform intellectual property negotiations. The general counsel or legal office of most institutions can provide guidance. The Internet site IFLA: Copyright and Intellectual Property Resources (see <http://www.nlc-bnc.ca/ifla/II/copyright.htm>) is a good resource for maintaining current awareness. It includes articles, reports and white papers, discussions, and information about organizations related to copyright issues, intellectual property in general, and electronic distribution of intellectual property.

THE INTELLECTUAL NATURE OF THE SOURCE MATERIALS

The following sections of this paper separately discuss the complement of considerations that bear on decisions to digitize. The elements are presented in a sequence that moves from relatively abstract assessments of intellectual value to nuts-and-bolts issues concerning whether available resources and technology can provide a product that meets expectations. In practice, the pieces interact in ways that are often complex.

Decisions about what to digitize must first and foremost address the intellectual value of the original sources. We are likely to be able to convert only a small percentage of existing scholarly materials to electronic form, and doing even this will require substantial investments. We therefore need to determine what it is truly worthwhile to convert.

Questions to Ask

Does the intellectual quality of the source material warrant the level of access made possible by digitizing?

Materials with marginal scholarly value are best left in their original form or made accessible in a less costly manner. Scholarly value, of course, is a subjective assessment and even the most marginal materials can support some kinds of research. Most users, nonetheless, would opt for electronic access to original monographs rather than to derivative works, or to the papers of a prominent scholar over the administrative records of a university department. Bibliographers regularly make purchase decisions that reflect their evaluation of the intellectual quality of single items or collections of materials. Similar judgments apply in choosing what to digitize.

Will digitization enhance the intellectual value of the material?

Scholarship can be facilitated when texts are made fully searchable by rekeying (retyping) them or by employing OCR software. Comparisons between successive drafts of a text and the final published work, for example, or with later editions and translations, are vastly simplified when the words and phrases are searchable. A concordance or thesaurus is likewise most easily mined when it is in searchable form. Electronic texts can be moved readily from one environment to another (from the World Wide Web onto the hard drive of a personal computer, and then into a word processing program, for example), shared with other users, and manipulated and reconfigured for multiple purposes. Digitized prints, drawings, and other visual resources can be viewed in groups at low resolution or inspected individually for very fine detail. Digital charts and tables, appropriately coded, can be loaded directly into statistical software packages for additional analysis. Census results, for instance, are

most easily used when the data have been formatted and imported into the Statistical Package for the Social Sciences.

Will electronic access to a body of information add significantly to its potential to enlighten, or are the original books, manuscripts, photographs, or paintings sufficient to the task?

A collection of thousands of portrait images, however promising a resource, might be nearly unapproachable because of its size and the condition and dimensions of individual items. Well-indexed and in digitized form, however, the collection could be searched with relative ease for images of a particular person or for some indexed characteristic (the country from which the portrait originates, for example). Likewise, the digitization of large-format architectural drawings could enable comparisons of small- and large-scale drawings, different views of the same architectural feature, or sequential phases of construction.

To what extent will the combination or aggregation of original sources increase their value?

Digitizing related scholarly monographs, like building a coherent collection of paper copies, can strengthen the context within which each title is approached. Ephemera—leaflets from a political campaign, for example—are often most useful when studied in the aggregate, as are posters, broadsides, and popular literature. Harvard has digitized daguerreotypes from thirteen repositories to facilitate the combinations and comparisons that are otherwise precluded by the fragility, value, and dispersion of the original images.

CURRENT AND POTENTIAL USERS

Some scholarly resources are heavily used; others are consulted infrequently. With only limited funds available for reformatting, types and levels of use can help to shape priorities.

Questions to Ask

Are scholars now consulting the proposed source materials? Are the materials being used as much as they might be?

These are complicated questions. Intensive use does not automatically make a collection a good candidate for digitizing. If the primary audience is local, for example, and if competition for a particular resource is not a problem, access may already be sufficient. Ephemera produced by a community political organization may be of great interest to local scholars and of limited value to a worldwide audience. On the other hand, if use is heavy and widespread, digitizing may at once guarantee convenient and reliable access, and make it possible for some institutions to discard their original copies. The JSTOR project (see <http://www.jstor.org/>), through which a large array of core scholarly journals is being made accessible in digital form, is a prime example of an initiative focusing on high-use materials.

Is current access to the proposed materials so difficult that digitization will create a new audience?

Low use may signal that a collection has marginal intellectual value, but there are many other reasons for valuable materials to have generated little interest. A collection may be held in a remote location, for example, or be owned by an institution with highly restrictive access policies. Bibliographic records may be poor, as is often the case with pamphlets. The value of digitizing such materials may go beyond the simple fact that the resulting files can be widely distributed. Broader access, as it creates a new community of users, can also facilitate more active scholarship.

Does the physical condition of the original materials limit their use?

Some resources are too fragile to be consulted. Aging newspapers or palm leaf manuscripts that break at the slightest flex simply cannot be browsed. In such cases, a digital copy might be provided to improve access, and a microfilm or other photographic surrogate made to ensure long-term survival. (Film can be made from a digital file or vice versa.)

Sources may also be at risk because of high user demand or extraordinary monetary value. A nation's founding documents, glass-plate

negatives of vanished architectural sites, or rare maps may benefit from the creation of digital copies that satisfy the purposes of most users. These files do not necessarily need to meet archival standards. They are created to protect the originals from handling.

Are related materials so widely dispersed that they cannot be studied in context?

Cooperative efforts to digitize disparate pieces of a greater whole can create or restore a more usable collection. Papyrus fragments, a prominent individual's far-flung correspondence, scattered photographs of a particular subject or by a specific photographer, and broken serial runs are among the many materials whose coherence, accessibility, and scholarly utility can be enhanced through digitization.

Will the proposed digital files be of manageable size and format?

Digital resources need to match users' technical capabilities and equipment. Most require Internet access and standard web browsers, or a CD-ROM drive. Images delivered to the Internet in formats other than JPEG or GIF require additional software for viewing or printing. Even when electronic resources are optimized for on-screen delivery, some network connections, particularly those via modem, are still far too slow to support browsing of digital collections at satisfactory speeds. And scholars in some locations may lack training opportunities or the ongoing technical support needed to take advantage of the electronic environment. These limitations, however, are not necessarily reasons to rule out digitizing. The worldwide trend is toward greater capabilities. Moreover, the more important the resources available electronically, the greater the incentive to acquire the network, viewing, and printing technology necessary to use those resources. Digitization may, in and of itself, stimulate improved access.

Will digitization address the needs of local students and scholars?

Immediate demand can inject a measure of practical reality into decisions to create electronic resources. An art historian might seek to scan art images and make them available to students as electronic reserves, as an alternative to slide-based classroom presentations and reviews. A historian may choose to teach from digitized images of manuscripts that would otherwise be unavailable to a large class. Because ready access to shared electronic files can transform the classroom, proposals to digitize in support of immediate teaching needs may garner faculty support.

ACTUAL AND ANTICIPATED NATURE OF USE

A person reading a book, looking at a photograph, or consulting a manuscript encounters few barriers to use. One might have to handle an object carefully, or use a magnifying glass to read fine print, but in general the work is immediately approachable. The same resource, when digitized, should be equally accessible and approachable. Ideally, the electronic version will also permit new kinds of use and more sophisticated types of analysis.

Questions to Ask

How do scholars use the existing source materials? What approach to digitization will facilitate their work?

Different digitizing techniques result in electronic files with different characteristics. These in turn can correspond well or poorly with scholarly needs. If the goal is to provide an image-based finding aid that helps users identify original materials of interest, for example, mounting slow-loading high-resolution images would be counterproductive. If, on the other hand, the intention is to reduce or eliminate handling of original materials, an image that fails to convey all critical information embodied in the original will fail to serve its intended purpose.

The simplest approach to digitizing involves use of a scanner or digital camera to create electronic pictures (bitmap images) of original materials. Decisions concerning the number of dots recorded by the scanner (resolution), how many shades of gray or colors will be recorded (bit depth), and other factors related to scanning equipment and settings will determine how well the digital product replicates the original. High-quality bitmap images can usually capture all the significant detail in texts or graphics. Scanning rare and unique texts or visual resources can make them accessible to users who would otherwise never see them. In such a case, merely reproducing the original in electronic form represents an extraordinary enhancement.

For textual materials, post-scan processing can support expanded capabilities. Scanned text can be processed with Optical Character Recognition software to produce searchable indexes. OCR software is now only occasionally employed in digitizing projects because it cannot yet interpret accurately all fonts and alphabets, and because it adds significantly to per-page costs. Text can also be rekeyed to create ASCII files—very straightforward digital text files that permit searching by keywords or phrases. In some cases this enhancement is the primary justification for digitization. Directories, dictionaries, and indexes are all significantly easier to use when specific words can be searched within a well-designed digital file.

ASCII texts accommodate key-word searching (e.g., searching for all instances of the word "temperance") and some kinds of analysis, but they do not readily replicate the structure and format of an original document. Without special coding, researchers cannot directly consult the seventh paragraph of the third chapter of a particular text. Nor can they search for all occurrences of "welcome" used as a verb rather than a noun. These capabilities become possible in marked-up texts, which are coded to highlight elements of structure, format, and syntax. The Standard Generalized Markup Language (SGML) is the emerging model. One SGML application, the Encoded Archival Description (EAD), is being used to create electronic versions of archival finding aids.

These and other approaches to digitizing carry very different costs, benefits, and resource requirements. While electronic versions can be more versatile than original materials, in some cases they hinder research. A scholar studying bookbindings or papermaking, for example, is poorly served by a reproduction of any kind. So too is the scholar whose immediate access to a large and important collection of literary works is sacrificed in order to serve a worldwide constituency—perhaps because bound volumes have been disbound for scanning.

Will digitization increase the utility of the source materials? Will it enable new kinds of teaching or research? Do scholars agree that the proposed product will be useful?

Digitization can enhance original materials in many ways. Image quality can be improved by eliminating extraneous stains and marks. Thumbnail images of visual resources (photographs, drawings, paintings) can be browsed to discover patterns, trends, and relationships among individual items, and specific images can then be scrutinized at higher resolution. Likewise, patrons can review scanned images to identify needed materials before requesting that they be retrieved from storage.

Electronic transcriptions of texts, in ASCII format or marked-up files, can be linked to bitmap images of original documents. Readers can then decide for themselves whether "authoritative" transcriptions are in fact accurate. Comparisons of different versions of a text are likewise simplified. Related texts and images can be assembled together within a single, unified corpus. Examples such as the Dante Project mounted by Dartmouth College (see <http://miltonsweb.mse.jhu.edu/dbases/dante.html>), which reproduces and links related texts and commentaries concerning the Divine Comedy; and Tufts University's Perseus Project (see <http://medusa.perseus.tufts.edu/>), an interactive, multimedia database on Archaic and Classical Greece, suggest the potential of electronic texts.

Almost all electronic products will provide basic links that allow users to navigate them (to locate a particular map within a printed text, for example). The degree to which a digitization project exploits electronic links will depend upon its intended use. For digital resources created as pedagogical tools, predetermined connections are part of the package. Products intended for research tend to be less aggressive in ordaining relationships among sources, since their creators assume that researchers will build their own structures of meaning.

Are there other scholars, librarians, and archivists who can collaborate to create a useful product?

Colleagues and potential users can clarify ideas, help select meaningful materials for conversion, improve project design, and stimulate early interest. "User demand" reflects both the intrinsic utility of specific source materials as well as a social context of participation and promotion.

THE FORMAT AND NATURE OF THE DIGITAL PRODUCT

Decisions to digitize must take into account the physical size, nature, and condition of source materials as they affect the characteristics of the desired product. They must likewise address whether available means of conversion can satisfy expectations for the result. Projects must also, from the very first, consider how users will be guided through the electronic version.

Questions to Ask

What critical features of the source material must be captured in the digital product? Are very high resolution copies, accurate rendition of colors, a seamless combination of images and text, or other qualities considered essential?

> The cost and nature of digitizing hardware and software continue to evolve, and preferred solutions are likely to shift as well. It may sometimes make sense to defer certain digitizing projects so that technology can catch up to needs. The success of a project to digitize oversized maps at Columbia University, for example, depended partly on the ability of users to see detail and read place names. As a result, the maps were scanned at relatively high resolution, thereby creating challenges for digital image delivery and presentation. File sizes were very large and initially outran the capacity of the library's computers and network. Greater bandwidth and more powerful machines have enhanced functionality.

If the original sources are to be retained, can they withstand the digitization process?

> Automatic sheet feeders are fast and efficient, but they may destroy brittle paper. Digital cameras can minimize the manipulation of source materials, but subjecting certain media—watercolors, for example—to prolonged lighting is problematic.

What type of hardware should be used for conversion?

> Color slides, for instance, cannot be fully represented by scanners that create only black-and-white images. Even a color scanner with limited capacity to reproduce tonalities will be inadequate when high-quality images are important. Digitizing equipment can be expensive, and the costs may be difficult to justify when use is sporadic. Some projects may thus be done most economically if they are contracted out. Agreements with external vendors, in addition to specifying technical conditions, performance expectations, and handling guidelines, must fully define ownership and distribution rights for all digital products.

Will a digitized sample meet users' needs? If so, how should the sample be constructed?

Many collections are too large to convert in their entirety. In the case of an artist's drawings, one might select materials from each of the artist's major periods, or representatives of the various media in which he or she worked, or particular subjects, such as cityscapes or portraits. Subsets of large collections can be defined in many ways and for many purposes. Collaboration with scholars and other experts is essential.

Will the information resources upon which the project is based continue to grow?

Ongoing commitments and extended arrangements for copyrights may be required when collections are still expanding, as is the case with current journals and annual reports, or the papers of a living individual. Consultations with scholars and other experts can be particularly useful, since the long-term value of current materials is often difficult to discern.

How will users navigate within and among digital collections?

Printed sources orient readers by means of tables of contents, chapters and sections, pagination, indexing, and formatting cues. Manuscript materials often rely on finding aids linked to the organization of file folders. Photographs may be mounted in albums. At a minimum, electronic products need to provide the same kind of functionality. The process may require several steps. For a multi-volume work that has been scanned page by page, for instance, each page is a separate computer file that must be individually labeled and stored. The files for critical pages of the work—for example, the title page, table of contents, and the first page of every new chapter—must then be linked to electronic navigational tools so that they can be easily located.

DESCRIBING, DELIVERING, AND RETAINING THE DIGITAL PRODUCT

While libraries can point with pride to their collective achievements in organizing and describing an enormous number and variety of collections and material types, some perplexing issues have not yet been resolved. There is still no consensus on how to handle ephemera that cannot realistically be cataloged by the piece and that are too insubstantial to shelve like most books and journals. Providing access to mixed media (a book accompanied by a floppy disk or CD-ROM, for example) is likewise problematic. But these issues, complicated as they are, pale next to the challenges of making digital products available to users. Decisions as to what resources should be digitized must be informed by an understanding of how the product will be described for users, delivered to them, and managed over time.

Questions to Ask

How will users know that the digital file exists?

Bibliographic records, finding aids, and indexes can all be adapted to include references to electronic resources. Nonetheless, our ability to determine what has been digitized remains well behind what we know about materials that have been microfilmed or photocopied.

One of the principal challenges is to determine what information is essential in describing an electronic product. The "Dublin Core" (see <http://purl.oclc.org/metadata/dublin_core/>) and other special initiatives for structuring and standardizing descriptive data propose to combine information about the technical characteristics of digital files (how they were created), their location, and a summary of their contents. The resulting records are known as "metadata." Their function is to provide users with a standardized means for intellectual access to digitized materials. Despite these and other initiatives, projects to catalog digital files are only in the developmental stage. No system has yet been widely adopted for tracking the digitizing activities of libraries, archives, and museums, although new approaches continue to emerge.

How can the digital product best be delivered to users?

Alternative modes of digital storage and delivery must be considered from the outset of a project. CD-ROMs, for instance, are distributed and used differently from information made accessible over the Internet. The differences are reflected in hardware requirements, software, and ease of use. CD-ROMs are sometimes bundled with software for searching and analysis that is superior to that generally provided for Internet files. On the other hand, access to CD-ROMs is limited to individual workstations or small networks, while Internet files can be made available to a very broad audience. And Internet

resources, by nature, can be updated or augmented without requiring users to replace objects that have become obsolete.

Internet products, however, generate questions of their own. How immediate must access be? Files can be mounted on a server so that they are instantaneously available on-line. They can be stored on disks in a jukebox and loaded on demand ("near-line" access), or kept off-site ("off-line") and retrieved and delivered on demand. Near-line and off-line access can save on server space and requirements, though there are countervailing staff costs associated with retrieving and mounting the files. Expected demand, file sizes, fee structures, and available staffing and equipment must all be considered.

Who will be authorized to use the digital resource, and under what circumstances?

Copyright holders may limit distribution rights, institutions may be unable or unwilling to provide the infrastructure needed to support universal access, and cost-recovery enterprises cannot by definition make their products available without restriction. Digitizing projects must thus consider access policies and control, pricing mechanisms, and billing procedures. Access issues impinge upon selection decisions in a number of ways. A university may mount high-resolution images of unique holdings for scholarly use (a medieval manuscript, an important collection of drawings), but would not allow unauthorized publication of those images. Moreover, electronic resources cost money that must be secured through subsidies or fees. When neither internal budgets nor external subventions provide adequate financial support, digitization will require a paying audience.

Access, when it is not universal, must be managed. Current alternatives include passwords, direct user fees, and limitations according to organizational affiliation. Different capabilities for viewing, downloading, and printing may be offered at different prices or to different sets of users. There are many options, each reflecting a different pathway toward a self-sustaining endeavor.

How will the integrity of the digitized data be ensured?

The malleability of electronic products makes them particularly useful for many kinds of scholarship. Digitized files must be embedded with detailed information concerning the methods used to create them. The same information should be included in external bibliographic or descriptive records. Users who are consulting or copying the sources must also be able to confirm that the files they see or receive match the originals. Means to authenticate and protect digital products, long available in financial and industrial applications, are only beginning to take hold in the scholarly world.

Particularly for digital products created to meet local demand, is the existing technology infrastructure adequate?

Robust computer systems and an appropriate number of work stations are perhaps more easily provided than such ancillary features as network printing capabilities in the library and in offices, classrooms, and residences.

What are the long-term intentions for the digital file?

In the case of electronic document delivery systems such as ARIEL (a product of the Research Libraries Group, Inc.), the goal in most cases is to provide very rapid access to specific articles or chapters. While images must be legible, they need not be perfect replicas; and copyright constraints, indexing complexities, and storage economies make it simpler to rescan on demand than to organize and retain random files. In other cases, the file may be kept for a longer, but still limited, period and then discarded—a reserve reading list or copyrighted images of artworks scanned to support classroom teaching, for example.

Is the long-term preservation of deteriorated materials a project goal?

Preserving documentary resources in electronic format presumes that, to the greatest extent possible, all the information contained in the original material has been captured completely and accurately. This requires careful attention to significant detail, whether the smallest text character on a page or all the shades and tones of blue and green in a seascape. Targets for resolution, grayscale, and rendition of color either exist or are being developed to ensure the needed detail and fidelity.

Digital preservation also requires a supporting organization and infrastructure dedicated to storing the electronic files and to migrating them to new formats and/or media as technologies change. Unless these capacities are all in place, digital files cannot be regarded as permanent. Creating an enduring digital preservation master file is a multidimensional task with long-term implications. Hybrid projects, in which digital files are complemented by copies on microfilm, alkaline paper, or some other stable medium, provide the insurance that exclusively electronic projects do not.

Digital processes meet preservation objectives without pretending to permanence. In the case of Spain's Archivo de Indias, for instance, low resolution grayscale images were prepared so that fragile original documents, some more than five hundred years old, could be spared the rigors of repeated consultation. The digital files, while they fall well short of capturing all the information in the originals, nonetheless fulfill a vital preservation function.

RELATIONSHIPS TO OTHER DIGITAL EFFORTS

Digitization, like other reformatting endeavors, takes place within a context larger than a single institution, discipline, or country. Selection decisions should be informed by both duplicative and complementary efforts.

Questions to Ask

Have the materials proposed for digitization already been converted to electronic form?

As we have seen, it can be difficult to determine whether a specific item has been digitized and by what means. If an electronic copy does exist, is it accurate, satisfactorily functional, and accessible? Does it take advantage of the capabilities of current technologies? If the existing product does not serve the intended purposes of the proposed project, a new version may be warranted.

Can cooperative digitization efforts bring together a cohesive body of material that would otherwise remain disassociated?

Standardized descriptors and a common approach to indexing and storage can allow dispersed materials to be combined in an amalgamated digital resource. The process involves institutional alliances as well as technological conventions. Different levels of participation and different expectations for returns may affect the result. If one institution provides the majority of materials for a digital project, for example, with many others completing the whole, the "lead" institution may claim special consideration or returns, requiring extra negotiation.

Successful projects to combine digital resources through a common system for organization and delivery suggest a new kind of model for collection building. Even in preservation microfilming, cooperative efforts to preserve a single title typically involve assembling dispersed materials at a central location for filming, or bringing together film prepared at various locations for splicing and duplication. The workflow of digital collection development can remain radically decentralized provided a robust infrastructure for collaboration is in place. The Research Libraries Group project, Studies in Scarlet: Marriage, Women, and the Law, 1815-1914, is a case in point (see <http://www.rlg.org/scarlet/sis.html>). Six U.S. libraries and one in Great Britain have scanned trial accounts, case law, statutes, treatises, and other materials related to the theme expressed in the project title. RLG established file naming conventions and other guidelines, designed the interface, and will serve the images—one of several models being explored for the creation of virtual collections. The conceptual kinship with traditional collection development is clear.

COSTS AND BENEFITS

Cost-benefit analysis assesses the relationship between functionality, demand, and expense. Limited resources and competing demands on organizational time and energy mean that the analysis must be rigorous and complete. The costs of creating electronic resources vary considerably. File size (and the associated storage needs) and processing requirements account for part of the differences, though labor requirements are even more important. Bitmap images in black and white are relatively inexpensive to produce and store. Grayscale images, currently capable of capturing up to 254 shades of gray plus black and white, are more costly; color images are the most costly of all. In each case, images with higher resolution result in larger digital files.

Accurate ASCII files of searchable text, even though occupying far less computer memory than any image file, are more expensive to produce than bitmaps of the same material. The main reason is that OCR software is not yet fully reliable. Materials converted by machine must be painstakingly proofread, or the source documents must be rekeyed in combination with careful attention to the detection and correction of errors. Costs rise even more for marked-up text, which entails yet another level of analysis and intervention. Creating other kinds of special databases or enhanced capabilities, for image files or for text, likewise raises the costs.

Costs vary even within specific approaches to digitization. All other things being equal, for example, it is less expensive to scan from single sheets than from bound volumes. Small sheets are less expensive to scan than oversized ones. Items in good condition are less costly to process than those that are deteriorated and thus require special handling.

Available cost figures for digitizing projects are often misleading. Cost projections seek to pin down a rapidly moving target. Although the prices of computer storage and processing power, for example, continue to fall, most projections simply extrapolate from available information about current price structures. Analyses often fail to account for certain categories of effort that, were they included, would alter cost calculations significantly. Labor expenses, for instance, often reflect only a pro-rated price per page that overlooks the real cost of a full-time employee. Crucial pieces of the workflow are sometimes written off as one-time "research and development" expenses. Functions such as preparation of materials for scanning, indexing, bibliographic description, post-scan processing, and long-term file management may not be factored into cost equations. Incomplete cost analyses can impute benefits that are difficult to represent on a project balance sheet. It may be true, for example, that ready access to backfiles of digitized journals will ultimately reduce or eliminate construction costs for new stack space. Unfortunately, money not spent on capital projects is unlikely to be reflected in support for other library initiatives. Though digitizing projects must calculate the likely costs and benefits, our ability to predict either of them is as yet rudimentary.

Questions to Ask

Who will benefit from the proposed digital product?

It is important to consider whether the product will support better teaching or research and enable students to learn more, or in different ways—if, for example, texts or images are more fully revealed. Digitization may allow librarians to manage collections and provide services more effectively, or to provide traditional services such as copying or interlibrary loan at lower cost or at less risk to collections.

Is the intellectual value of the proposed product commensurate with the expense?

The limited resources available for digitization might have greater impact if they were directed at another project, or directed toward an entirely different approach to providing access—through exhaustive indexing perhaps, or microfilming, or some other type of reformatting that would prove in the end more useful to scholars.

Could an acceptable product be created at lower cost?

When materials are scanned to support short-term course work, for example, careful (and expensive) post-scan processing to eliminate extraneous marks and speckles or to deskew misaligned images may be a waste of time. Likewise, an adequate substitute for full-text scanning of little-used journals might be provided by linking scanned tables of contents and indexes to bibliographic records and relying on traditional forms of document delivery.

How will the proposed project address the long-term costs associated with digital files?

The accumulated body of digital products may enable savings elsewhere in the institution—for example, by reducing staff costs for reshelving bound journals, or by lowering the costs of storage, circulation, and preservation—and these savings could offset some or all of the expense of digitizing. But such savings as may be realized are difficult to predict. It is essential to realize that the costs of digitization are just beginning at the time of initial capture. The programmatic capacity to distribute and maintain electronic resources, and to migrate them to new forms as original digital platforms fail and formats and software are superseded, is fundamental to long-term efforts. In addition, there are staff costs associated with training and user support. Finally, rising user expectations may require that existing digital files be reprocessed in new ways. When OCR software is perfected, for example, unsearchable bitmap images of texts could be thought unsatisfactory. Projects that do not plan for change may become obsolete, and therefore irrelevant.

Can external funding be secured to support the proposed project?

Some foundations are particularly interested in electronic products, and specialized scholarly initiatives may attract their support.

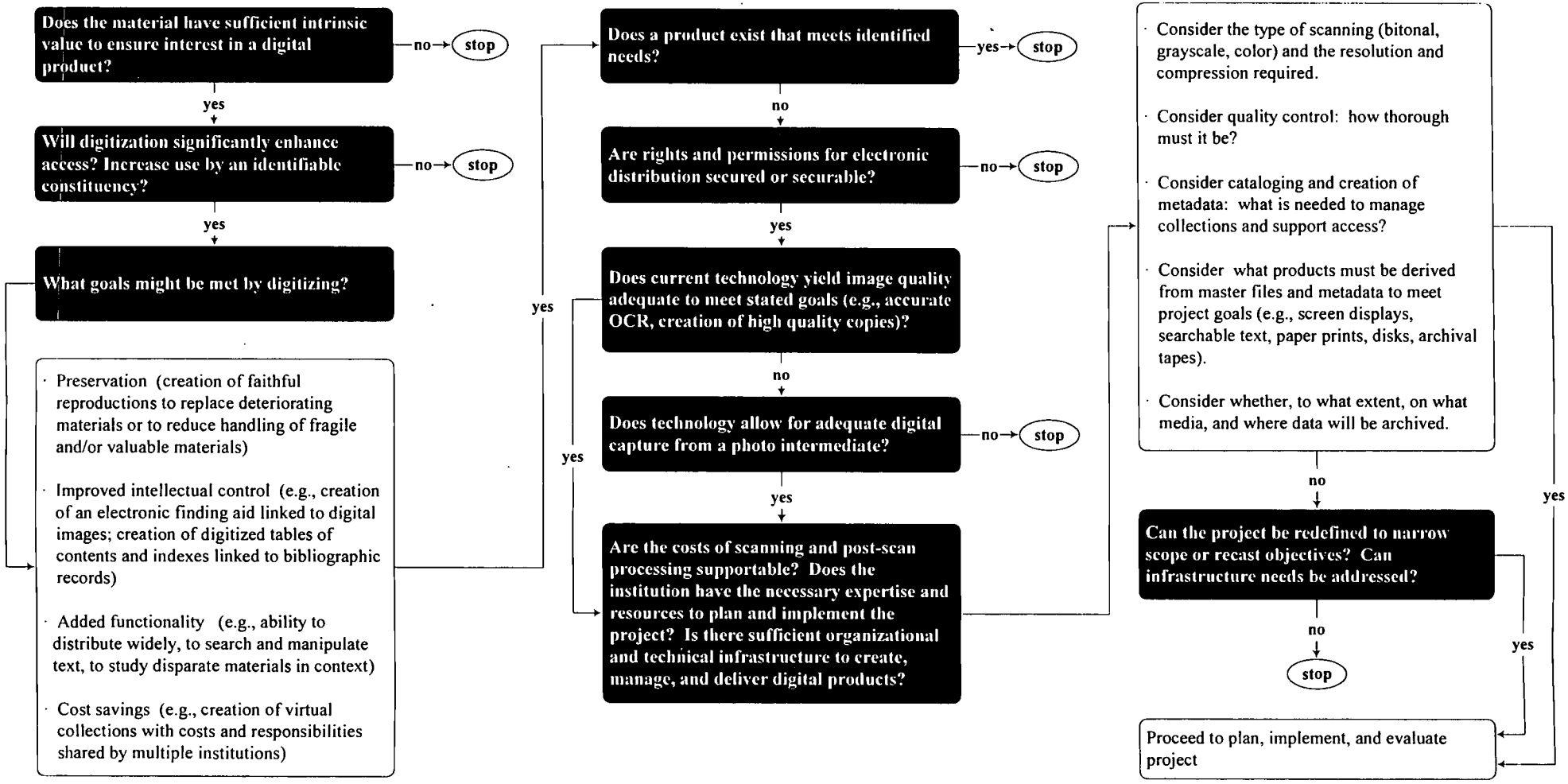
CONCLUSION

Research libraries are eagerly embracing the digital world. They are acquiring access to great quantities of electronic materials produced outside their walls and are making digital versions of their own holdings. These projects, as they become more common, are bringing both the broad issues and the nuances of the digitizing process into sharper relief.

Projects based on careful review, analysis, and planning can yield electronic resources that are functional and faithful to the original sources, and that support new kinds of scholarship. A detailed plan of work, regular assessment of progress, closely documented adjustments and corrections, and the retention of other project-related data can strengthen the knowledge base for future efforts. Each success, as well as each failure, will bring us closer to fulfilling the promises of the electronic environment.

The process of deciding what to digitize anticipates all the major stages of project implementation. Digital resources depend on the nature and importance of the original source materials, but also on the nature and quality of the digitizing process itself—on how well relevant information is captured from the original, and then on how the digital data are organized, indexed, delivered to users, and maintained over time. Disciplined efforts to address the themes and questions outlined in this essay will help ensure that new digitizing projects fulfill the expectations of libraries, students, and scholars.

SELECTION FOR DIGITIZING:
A Decision-Making Matrix



SChapman, DHazen, JHorrell, JMerrill-Oldham

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").