What to Withdraw?
Print Collections Management in the Wake of Digitization

September 29, 2009

Authors:
Roger C. Schonfeld (Manager of Research) & Ross Housewright (Analyst)

Ithaka S+R (www.ithaka.org/ithaka-s-r) is the strategy and research arm of ITHAKA, a not-for-profit organization dedicated to helping the academic community use digital technologies to preserve the scholarly record and to advance research and teaching in sustainable ways. The Ithaka S+R team supports innovation in higher education by working with initiatives and organizations to develop sustainable business models and by conducting research and analysis on the impact of digital media on the academic community as a whole. Insights from these efforts are shared broadly, with more than a dozen reports freely available online. JSTOR, an accessible archive of more than 1,000 scholarly journals and other content, and Portico, a service that preserves content published in electronic form for future generations, are also part of ITHAKA.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The large-scale digitization of print journal collections has led to most access needs being met via digital surrogates. Numerous libraries would therefore like to reassign the space occupied by print collections towards higher-value uses. To aid their planning, this report addresses two key questions: which types of print journals can libraries withdraw responsibly today, and how can that set of materials be expanded to allow libraries the maximum possible flexibility?

For those journals where print no longer serves an important access role, preservation is the format's principal remaining role. The study therefore undertakes a system-wide analysis of the purpose of retaining print for preservation purposes, looking at the needs of all libraries and their users collectively.

This analysis finds several rationales for retaining some copies of the print version: the need to fix scanning errors; insufficient reliability of the digital provider; inadequate preservation of the digitized versions; the presence of significant quantities of important non-textual material that may be poorly represented in digital form; and campus political considerations. The appropriate disposition of print copies of a given journal should vary depending on the characteristics of the print original and its digitized version in each of these categories.

Because many of the rationales for retaining print are likely to decline over the course of time, this report introduces time horizons for print preservation. Librarians have often discussed preservation responsibilities as if it were possible to undertake perpetual commitments, but specified time commitments coupled with regular reassessment of priorities and responsibilities permit better decision-making. The model we propose therefore examines the minimum period of time that access will be needed to at least one copy of the print original.

While complex, this methodology provides for preservation frameworks that vary based on risk profiles. For example, text-only materials require less concern than image-intensive materials, while high-quality digitization processes digital preservation practices similarly indicate lower concern. These rationales indicate the need for at least one print copy of well-digitized digitally preserved text-only materials to be available for at least 20 years.

In order to guard against losses over time and assure the availability of a single copy after the stated time horizon, a greater number of print copies of any digitized title need to be secured today. In the exemplar scenario, a minimum of two page-verified print repository copies would be needed.

When such well-digitized digitally-preserved text-only journals are held in two page-verified print repository environments, therefore, other libraries can safely withdraw their print holdings if they so choose.

Most journals do not meet the criteria for withdrawal, so the report therefore provides several strategies to expand the set of materials that meet these criteria. First, organizations pursuing digitization projects should more transparent about their standards and practices. Second, when digitization quality is low, it should be upgraded over the course of time. Finally, the library community should aggregate the work of existing mechanisms for print storage, de-duplication, and preservation, so that print repositories can more effectively contribute to a system-wide withdrawals strategy.

INTRODUCTION

In recent years, libraries have licensed access to the digitized versions of numerous journal backfiles. Digitization has expanded access to and use of these backfiles tremendously, and libraries see an opportunity to reassign the space occupied by their rarely used print counterparts to higher priority purposes. Some libraries are already withdrawing print versions, while others are actively contemplating the possibility. Without coordination, some libraries will inevitably withdraw print holdings in a way that unintentionally removes certain materials altogether from community-wide print holdings. Which types of materials can be withdrawn responsibly today and how can that set of materials be expanded to allow libraries the maximum possible flexibility? The purpose of this report is to address these two fundamental questions about print collection management in the digitized environment.

Our intention in this report is to contribute to the assurance of the preservation of intellectual content. We recognize that print withdrawals are becoming increasingly widespread at libraries that previously would have taken an active role in print preservation, and only limited resources will realistically be devoted to preserving an outmoded format for these materials. It is therefore of the highest priority that we identify opportunities where libraries can achieve increased flexibility, so that their limited print-format resources can be allocated in a targeted and strategic way.[1]

While other work has addressed these topics, this report takes a completely different perspective:

- First, and most importantly, our analysis proceeds from a system-wide perspective, recognizing that although the aggregated system-wide value of maintaining a print artifact may be significant, the value of doing so to an individual library or small group of them is likely out of step with the costs.
- Second, we do not assume that there is any intrinsic value to the maintenance of collections of print artifacts but rather take a critical perspective to analyze why the community might want to keep any print at all.
- Third, although there may be rationales for retaining some print copies, we do not assume that these print copies should be retained forever, but rather that minimum time horizons should be established for such retention.

The questions about withdrawing print versions that this report considers are appropriate for general collections of published materials – and inappropriate for rare and unique collections. We focus our attention on scholarly journals, where the greatest amount of research and policy analysis has been conducted to date and where the digital transition has progressed furthest. Still, widespread print withdrawals have already taken place for other content types such as reference resources, and the rapidly expanding mass digitization programs suggest that similar collection management questions will increasingly apply to monographs, government documents, and other materials as well.

We draw numerous examples from JSTOR's experience digitizing journals, the oldest effort to digitize journal backfiles and the one with which we have the greatest familiarity, both due to research studies we have conducted as well as because of our shared organizational home at Ithaka. We focus exclusively on journals that JSTOR has digitized itself, rather than on other content types on the JSTOR platform or materials produced with partners. Although we frequently draw on examples about the JSTOR-digitized

[1] It is important to emphasize that non-research institutions that have always actively managed their collections should not grow more conservative in their collection management practices because of digitization. Rather, this report focuses on when and how those libraries that have traditionally played an important role in retention and preservation of print general collections can responsibly transition such preservation responsibilities to the digital format and the system-wide level.

journals, we emphasize to the greatest extent possible the limits of that experience as well as examples drawn from other sources.

THE VALUE OF CAMPUS PRINT COLLECTIONS

Before robust interlibrary sharing networks were developed, acquiring physical copies of scholarly materials in a local library in anticipation of future demand was by far the most effective way to serve the needs of local users. As a result, although collections were developed largely independently of one another, significant overlap developed between library collections, with many scholarly materials being held in parallel at multiple institutions for local use. Indeed, as new colleges and universities were developed, overlap increased further. Some libraries might deaccession some of their holdings based on space constraints or other considerations, but it was widely assumed that materials would remain available somewhere in the system due to this overlap. The need for local access drove acquisitions and retention choices, which aggregated across the system to unintentionally but effectively yield substantial overlap. Many libraries managed collections actively, weeding regularly, but others built historically strong research collections by de-accessioning rarely and only under special circumstances. A side effect of substantial overlap between library collections combined with long-term retention was the effective accomplishment of many community preservation goals, but such preservation was a happy by-product rather than a planned outcome. Historically, such overlap was the only way in which local access to necessary materials – a primary function of the library – could be effectively provided.[2]

Nevertheless, since the turn of the 20th century, library leaders have recognized the inefficiency of overlapping collections developed exclusively based on local needs and proposed various collaborations that would decrease overlap in local collections while maintaining *access* to collections through shared ownership of materials. Still, given the challenges in sharing information about collections and coordinating collection development, the decentralized "just in case" model prevailed for most of the 20th Century.[3]

With the development of robust interlibrary sharing networks and remotely accessible OPACs in the 1970s and 1980s, it became possible for libraries to make collecting and retention choices with an awareness of peer library behavior. Libraries could ensure access to materials of interest without maintaining them in local collections, instead relying increasingly on the ability to request materials of interest via interlibrary loan.[4] As libraries came to consider their local collections in this context, interest in crafting more efficient mechanisms to coordinate collections developed rapidly. For example, some groups of libraries, often at the state level, attempted to develop "last copy" policies to govern deaccession practices.[5] Perhaps the most significant example was the vision for RLG to coordinate research library collecting at a national level.[6] Such initiatives demonstrated the substantial interest of even many of the largest research libraries, when faced with budgetary challenges, in coordinating their work to achieve greater efficiencies.

[2] This issue is explored in greater depth in Roger Schonfeld, "Commodity Collections: The Role of American Academic Libraries in the Maintenance of Knowledge, 1876-1900" (presented at the Society for the History of Authorship, Reading, and Publishing, Halifax, Nova Scotia, Canada, June 2005). Quantitative analysis of the growth in overlap in book collections may be found in Roger C. Schonfeld and Brian F. Lavoie, "Books without Boundaries: A Brief Tour of the System-wide Print Book Collection," *Journal of Electronic Publishing* 9, no. 2 (Summer 2006)

[3] See Roger Schonfeld, "The Role of Information-Sharing in Book Survivability in the United States, 1890-1940" (presented at the Society for the History of Technology, Las Vegas, NV, October 2006).

[4] See for example Scott Bennett, *Report on the Conoco Project in German Literature and Geology* (Mountain View, CA: Research Libraries Group, 1987) and T. Macky, "Interlibrary Loan: An Acceptable Alternative to Purchase," *Wilson Library Bulletin* 63, no. 5 (January 1989): 54-56.

[5] See for example Terry L Weech, Bryce L. Allen, and F. Wilfrid Lancaster, "The Last Copy Center Study. Illinois State Library Feasibility Study.," *Illinois Libraries*, no. 72 (1990): 44-50; "North Dakota Last Copy Retention Policy," *The Unabashed Librarian*, no. 83 (1992): 24.

[6] See for example Nancy E. Gwinn and Paul H. Mosher, "Coordinating Collection Development: The RLG Conspectus," *College and Research Libraries* 44, no. 2 (March 1983): 128-140.

Notwithstanding deep-seated interest in finding such efficiencies, technology and organizational issues limited what could be achieved in the United States. The political culture of the UK, for example, eventually allowed for much greater central coordination of collecting and preservation. But the United States has retained a fairly decentralized higher education system. In keeping with this decentralized tradition, there have been few coordinated efforts to manage collections at a system-wide level to accomplish community preservation goals. These have principally occurred for government information: the Federal Depository Libraries Program coordinated by the Government Printing Office, the National Network of Libraries of Medicine coordinated by the National Library of Medicine, and a number of similar federal initiatives. For most general collections, however, naturally occurring overlap across libraries, without any coordination, has as a byproduct provided the principal framework for preservation.

On a bilateral or consortial basis, numerous initiatives have attempted to coordinate collecting and preservation of print holdings. But deeper coordination has been more difficult to develop and sustain. In some few cases, preservation challenges that clearly could not be addressed through redundancy and overlap have brought about more coordinated action such as the Brittle Books Program and the Commission for Preservation and Access.[7] These have generally taken the form of coordinating initiatives, providing structure, guidance, and often funding for individual libraries to work towards a common end. They have addressed problems for discrete collection areas with some success through large-scale reformatting initiatives, though they have received criticism in some cases for providing an inadequate substitute for the original artifacts.[8] These initiatives did not attempt to shift organizational responsibility for preservation in any significant way.

Today, with a far more widespread and rapid reformatting in progress, fundamental questions that threaten the institutionally-organized overlap-driven model for print preservation are being raised about the value of campus print collections. Major publishers, JSTOR, Google, and others have digitized millions of books and journals, and much scholarship is likely to be communicated exclusively in digital form in the future. Faculty members have responded strongly, indicating a preference for e-journals over print that has only continued to grow.[9] And experimental research demonstrates that, in practice, faculty members have little ongoing need for access to print versions of journals.[10] The technology constraints that demanded local access to print have been largely obviated, in turn upending the overlap-driven print strategy.

The library community has moved into an environment in which print versions of journals go largely untouched in favor of digital versions that, when available, meet virtually all user needs, while competing needs for library space have continued to mount. As a result, the value to an institution of maintaining local physical journal collections has in many cases fallen precipitously. The direct costs of doing so, however, have not changed significantly, and new high priority needs for space have driven up the

[7] See, for example, Abby Smith, "The Future of the Past: Preservation in American Research Libraries" (Council on Library and Information Resources, April 1999)

[8] See Nicholson Baker, *Double Fold* (New York, NY: Random House, 2001) for some of the most strident criticism. For some of the largely healthy debate that this work stimulated, see Richard J. Cox, *Vandals in the Stacks? A Response to Nicholson Baker's Assault on Libraries*, vol. 98, Contributions in Librarianship and Information Science (Westport, Connecticut: Greenwood Press, 2002).

[9] Ithaka's 2006 faculty study shows that while most faculty believe it is important that print copies of journals be maintained somewhere in the library system, far fewer feel that such collections need to be preserved locally. This is most strongly seen among scientists, where only about a third felt that local print collections were necessary given the existence of digital collections. It is worth noting, however, that scholars in other disciplines – especially in the humanities – are significantly less comfortable with the switch to digital. Ross Housewright and Roger Schonfeld, *Ithaka's 2006 Studies of Key Stakeholders in the Digital Transformation in Higher Education* (Ithaka, August 18, 2008)

[10] Brian E.C. Schottlaender et al., "Collection Management Strategies in a Digital Environment," http://www.ucop.edu/cmi/finalreport/index.html.

opportunity cost of devoting high value space to the relatively low value activity of journal back issue storage.[11]

Given the increasing disparity between the cost and the value of maintaining local print journal backfile collections, the large-scale deaccessioning of these materials has become an increasingly rational choice when viewed from the perspective of any individual library. At long last, libraries can, if they so choose, eliminate local storage of collections while continuing to provide for reader needs.[12]

LIBRARY DECISION-MAKING

Thus, most libraries have a strong incentive to deaccession print journal backfiles available to them electronically. And this incentive is keenly felt. In 2006, over 40% of collection development directors at major research libraries agreed strongly that "in the near future, it will no longer be necessary for our library to maintain hard-copy versions of journals," while the share that believes print preservation is important stands to decline.[13] Library decision-making on the print to digital transition has been highly fragmented, typically not coordinated inter-institutionally. While some efforts have been made to bring strategic planning to the community, they have not gained significant traction.[14] This section, based on interviews with librarians and other key decision-makers from 14 colleges and universities, examines the institutional nature of library decision-making about the print to electronic transition.[15]

What actually motivates libraries to deaccession print journals? Some observers have suggested that by reducing the amount of space devoted to collections storage, libraries save the capital costs of facilities expansion. But libraries face several impediments to realizing the actual cost savings associated with space reductions. Most importantly, there is rarely a mechanism for libraries to be credited for cost savings as a result of careful space planning that obviates the need for facilities expansion, since they are realized to a capital budget that is rarely controlled by the library itself, which is responsible for only operating budgets. The two accounts are generally not fungible, and although creative budgeting can allow cost reductions in the capital budget to result in a credit to a libraries' operating budget, this is rarely done. Consequently, although in theory libraries might use estimates of cost-savings to prioritize space-savings initiatives as against other cost reductions, in practice such planning does not take place.

Instead, libraries look to available space, and the competing uses for space, as if it were a completely different account. Therefore, unless there is something they would prefer to do with the space near-term, they are unlikely to withdraw print materials proactively. Although all libraries face space pressures eventually, the timing of this pressure varies from library to library. The oscillation between pressure and

[11] The non-subscription costs of maintaining print collections far outstrip those associated with the maintenance of electronic collections. See Roger C. Schonfeld et al., *The Nonsubscription Side of Periodicals: Changes in Library Operations and Costs between Print and Electronic Formats* (Washington, D.C.: Council on Library and Information Resources, June 2004) And in recent years, many libraries have faced new priorities for space, such as a growing interest in supporting teaching and learning activities through the development of an information commons. See, for example, Katherine S. Mangan, "Packing Up the Books," *The Chronicle of Higher Education*, July 1, 2005

[12] For analysis of the largely separate problem of the transition of journal current issue subscriptions to electronic format, see Chandra Prabha, "Shifting from Print to Electronic Journals in ARL University Libraries," *Serials Review* 33, no. 1 (March 2007): 4-13

[13] This finding is from Ithaka's 2006 survey of librarians, which also demonstrates that while increasing attention is devoted to the need for digital preservation, this is in many ways supplanting interest in print preservation. Although almost 70% of librarians believed in 2006 that print journal preservation is currently an important function of the library, only about 50% believed it would be in five years from then. Conversely, while about the same number believed that electronic preservation is currently an important function, about 85% believed it would be in five years. Housewright and Schonfeld, *Ithaka's 2006 Studies of Key Stakeholders in the Digital Transformation in Higher Education*

[14] The Cornell-spearheaded Janus initiative, which perhaps showed the greatest promise, raised important issues but does not appear to have taken hold. See "Janus Conference on Research Library Collections, Cornell University, October 9-11,2005," http://www.library.cornell.edu/janusconference/index.html

[15] Interviews were conducted with librarians, faculty, students, and provosts across 14 institutions in 2007: Albion College, University of Arizona, Bellarmine University, Berea College, University of Chicago, Duquesne University, East Texas Baptist University, Elmhurst College, George Mason University, University of Louisville, University of Michigan, University of Minnesota, College of St. Benedict's/St. John's University, and Vanderbilt University.

relief that any individual library may face complicates community-wide strategic planning and investment.

A library may contemplate the deaccessioning of backfiles because it is out of space entirely or because it would like to utilize the space for a more highly-valued purpose. The information/learning commons movement to create suitable learning spaces and bring new services into the library has been transforming the physical space of library after library. As libraries seek an expanded role in the teaching and learning process, space is generally seen as a critical asset. When such a priority stands some chance of being funded, libraries turn to the deaccessioning of print as a key tactic for finding the needed space.

Once a library has reached the conviction that it will withdraw some print holdings, it typically has looked first to reference materials or periodicals. Many librarians do not view reference works as having any significant preservation considerations – at most libraries they are seen as superseded by new editions – thereby minimizing the complexity of withdrawing print versions when more useful electronic versions become available. But for periodicals and especially scholarly journals, preservation and perpetual access considerations are of significant importance at many libraries.

The most important source of backfile withdrawals at the 14 sampled institutions was from the journals that JSTOR has digitized. As JSTOR-digitized journals have been moved offsite or withdrawn outright, however, libraries have begun to look elsewhere: especially to the science collections, from the major commercial publishers that have undertaken backfile digitization programs, and that go completely unused in physical form. Where assurance of long-term access are provided, some libraries have been willing to withdraw print volumes of science journals from their collections. Still, libraries have tended to be more risk-averse in dealing with backfiles of commercial publishers as compared with JSTOR. As digitized versions of this content are deposited in trusted electronic archives, however, some libraries may feel more comfortable withdrawing the associated print versions. In both of these cases, the page-imaging, cover-to-cover scanning practices are important considerations for most libraries, which would be reluctant to withdrawn print versions if only the text, but not the images, advertisements, or front and back matter, were available digitally.

Libraries have only withdrawn print backfiles when they feel that there is adequate reason to believe in the long-term reasonable availability of the electronic versions, whether it is due to trust arising from the scanning practices, ownership-like licensing terms, or the not-for-profit status of a community-based resource. While the exact threshold varies from institution to institution, certain general principles emerge. For example, libraries tend not to withdraw backfile volumes that are available online only via an aggregator resource, because they do not believe that they have sufficient assurance of the reliability of their contents. These types of considerations are absolutely imperative, because in electing to rely on a digital version libraries can risk placing themselves at the mercy of the financial health of the vendor and its commitment to a reasonable pricing model.

In taking the decision to withdraw local print backfiles, most librarians nevertheless expect to see print versions remain available to their community from some remote location. Often, these expectations are based on informal and sometimes inaccurate assumptions. For example, many small colleges have long seen preservation as the purview of research universities, and have made local deaccessioning decisions on the assumption that other libraries have made a commitment to maintain print collections in perpetuity.

But many research universities are actively deaccessioning print materials. Although some few major research libraries may have formally committed to maintaining their existing print collections, information about these policies is difficult to obtain and hard to compare. Most libraries' long-term

intentions are not settled, making it very difficult for their peers to evaluate what collections will be maintained. Relevant collection management policies, where they exist, are often insufficient, promising to attempt to maintain materials only until local needs change. And even though these statements may be well-intentioned and strongly felt, they do not constitute binding obligations. Some libraries have created bilateral or multilateral agreements about the disposal of last copies, generally tied to a shared print repository, but these are generally also not transparent and non-binding beyond the group of libraries involved.

Informal arrangements, based on unspoken assumptions about the behaviors of others in the community, are clearly inadequate relative to expectations. There is a very real risk that so many copies may be discarded as to threaten the availability of certain materials in their original format. But should this inhibit individual libraries' withdrawals of print versions? In order to understand which materials are adequately safeguarded across the system such that withdrawals can safely proceed at a local level, we must first understand why the community might concern itself with the retention of print at all.

RATIONALES FOR COMMUNITY ATTENTION TO PRINT PRESERVATION

Many librarians argue that the community should retain and preserve at least some minimal number of print copies, so that print is not lost altogether. In conversation, many librarians suggest that five to ten copies seems about right. But the rationales given for such preservation are more often philosophical than practical.

What is the value of preserving not only the intellectual content but specifically the physical instantiation – the artifact itself? If the print has no remaining value, then libraries of all types could feel comfortable withdrawing their print copies without a second thought. If the print has a very high remaining value, then few if any libraries would consider withdrawing any print versions at all. Ideally, we would quantify the value and measure it against cost to determine how libraries should proceed. Given the challenges of doing so, we have made the pragmatic choice to articulate the sources of value rather than actually quantifying them.

Some – most notably Nicholson Baker – have advanced arguments against any deaccessioning of print materials. Inspired by his disappointment in the widespread replacement of print newspapers backfiles with inadequate microfilm surrogates, Baker argues, at least implicitly, that all print materials should be maintained in perpetuity. He suggests that surrogates fail to accurately reproduce printed materials, and that maintaining print collections as they have been is the best way to guard against any losses.[16] Although some of Baker's concerns deserve further contemplation, his implied preference for the perpetual retention of all print materials cannot possibly be feasible when libraries hold hundreds or thousands of ill-used copies, far beyond the number required for access or preservation purposes. A strategic approach that recognizes that many print materials will be deaccessioned may more realistically accomplish the community's practical preservation goals.

In contrast to Baker, Stephen Nichols and Abby Smith offer a strategic approach to print preservation aimed at addressing a more narrowly-defined problem. Their concern is the maintenance of access to artifacts, which they thoughtfully define as "things that have intrinsic value as objects, independent of

[16] Baker's argument unfolds indirectly, but he states clearly that "all microfilming and digital scanning [should] be nondestructive," implying that no library should ever withdraw print originals. Baker, *Double Fold*, 270.

their informational content."[17] Nichols and Smith distinguished different classes of materials with varying artifactual value. One example in the realm of journals might be periodicals printed on hand presses, or before the dawn of modern mechanical typesetting and printing technologies in the early 20th century, which can have significant object-level differences. But modern publishing technologies do not result in such individual distinctions. In the absence of important marginalia or other rare and unique features, therefore, identical print journals are interchangeable, eliminating the need to consider each one as an individual artifact. Nichols and Smith developed strategies that varied depending on the individual characteristics and projected uses of the different classes of materials in question. They anticipated the need to determine with certainty what "last, best" copies remained of an artifact of published materials and assign responsibility for the stewardship of such items.

In this report, we similarly hope to define a clear set of community preservation goals and use them to develop a strategy for print withdrawals that ensures preservation. For this reason, we believe it is imperative that the identification of these goals include voices from the rare books, special collections, archives, and preservation communities, along with the collection management community, and also from library and academic leaders obliged to make difficult resource-prioritization choices with constrained resources. In this section, we identify a set of community goals for the preservation of print originals in an effort to initiate a dispassionate discussion about the community's practical preservation needs. A healthy debate on this topic is important, and others may come to different conclusions regarding these system requirements.

Fix scanning errors

We look to the digitization and access lifecycle for print retention requirements, first considering the problem of shortcomings in the scanning process. Errors such as a blurred scan, an obscured image, or a skipped page, result in a digital object which is not a perfect – often not even an adequate – surrogate for a print original. Although stringent quality control procedures during initial digitization can minimize the need for such redigitization, even very strict processes will let through some small number of errors.

For example, although JSTOR has utilized an extremely quality-oriented approach to digitization, as a part of which each page is reviewed by a human prior to scanning to ensure that nothing is missing from the original source, it has had to return to print originals to correct hundreds of such errors. But while some digitization initiatives may aim for rigorous quality, many others fundamentally lack this emphasis on preservation quality, instead prioritizing speed to enable access to a wider range of materials. Google Books is an example of such an explicitly access-oriented digitization program, and while it has successfully made a substantial number of books available digitally in a short period of time, the correspondingly high rate of errors has been noted by many.[18]

Unfortunately, the line between access- and preservation-oriented digitization projects is often blurred by observers projecting their own concerns on initiatives that are unclear about their ultimate goals, and digital collections that never intended to serve as preservation reformatting efforts may be inappropriately

[17] Stephen G. Nichols and Abby Smith, "The Evidence in Hand: Report of the Task Force on the Artifact in Library Collections" (Council on Library and Information Resources, 2001), http://www.clir.org/pubs/reports/pub103/contents.html.

[18] Duguid presents an overview of some of the various ways in which Google Books may fail to live up to the quality expectations of scholars in Paul Duguid, "Inheritance and loss? A brief survey of Google Books," *First Monday* 12, no. 8 (August 2007), http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1972/1847 And many others have anecdotally noted holes and errors in Google's digitization processes that emphasize its nature as an *access* rather than a *preservation* tool – see, for example, Dan Cohen, "Google Fingers," June 26, 2006, http://www.dancohen.org/blog/posts/google_fingers. And in an 8/14/09 interview, Kurt Groetcsh, Technical Collections Specialist, and Ben Bunnell, Manager, Library Partnerships Team of Google Book Search expressed concern to the authors with the idea that libraries might choose to deaccession print materials based on their digitization as a part of this program, stating that Google's digitization "is an access project," and that they "are not trying to pretend to be a preservation process," Although these may be acceptable compromises in the production of a large-scale access-oriented digitization program, necessary to achieve scale, the resulting archive is not appropriate for preservation needs.

relied upon. Greater transparency is needed about the quality standards used by various digitization programs, so that the community can better determine the level to which digitized collections can be relied upon as accurate. But even in the case of preservation-oriented digitization efforts with stringent up-front quality controls, the lack of a print collection that can serve as a source for redigitization as needed entails a degree of risk of permanent information loss.

Still, the problem of insufficient digitization quality can be seen as self-correcting. If users are provided with an easy way to inform the digitizer of errors noticed in a collection, and if the digitizer is responsive and accurately corrects these errors in a timely fashion, then over time most errors may be identified and addressed – a crowd-sourced quality assurance program. Such a post-digitization verification process could be used to flag and correct collections scanned primarily for access, in a sense "promoting" a set of materials that have been reviewed and deemed acceptable into a preservation-oriented digital collection. It takes time for the majority of pages in a set of digitized journal backfiles to be viewed by enough users that it is likely that an error, if present, would be reported. For example, half of the error reports JSTOR has received have come within two years of online availability, while 92% come within five years. While crowd-sourcing cannot identify all problems, it may be reasonable to assume that those errors not noticed and reported over a reasonable period of time are associated with articles of relatively low value due to their non-use. Beyond a certain point, the costs of preserving an entire set of print materials against the eventuality that an additional error in a minimally used corner of the collection will be discovered grow untenable. Consequently, if a provider is responsible in its correction of user-reported errors, the need to retain print for such purposes may grow minimal after a sufficient amount of time has passed after materials were made available digitally. If print version can be securely retained long enough that the overwhelming majority of errors are caught, risks of information loss can be kept to an acceptable level and the costs of print preservation can be contained and anticipated.[19]

This section has made clear the wide disparity in initial scanning practices and subsequent error-correction approaches. Consequently, the time horizon for print retention to support these needs may vary significantly.

Inadequate scanning standards / practices

Second, even if materials have been accurately digitized at high levels of quality, the ability to re-digitize from print originals may remain valuable if new goals arise that cannot be achieved with the first round of scanning. For example, JSTOR's initial digitization processes created bitonal scans of all journal pages and, when appropriate, full-color scans of images that were linked from the journal page. When JSTOR sought to produce a single composite page image that offered the best of both worlds – more readable black and white text alongside full-color images – it found that these two scans could often not be accurately combined and that new scans were necessary to effectively produce a composite image. And elsewhere, some digitizers have found that for certain classes of materials, new optical character recognition technologies can offer greater accuracy if applied to a grayscale scan rather than a bitonal, so increasing accuracy in this way might require re-digitization from print originals. Many needs for redigitization may be obviated through proactive consideration of the uses for the digital surrogates, digitization at generously high levels of quality, and the retention of raw files. But, large-scale scanning efforts inevitably face tradeoffs between quality and cost, and these choices unquestionably limit what

[19] The notion of "selection for preservation" – the fact that "there are more library books and journals in need of preservation today than can possibly be saved before they crumble and disappear," and that limited resources require choices about preservation priorities – has long been recognized by the preservation community. See, for example, The Commission on Preservation and Access, "Selection for Preservation of Research Library Materials," August 1989, http://www.clir.org/pubs/reports/lesk/select.html

can be done with the resulting digital file: greater detail cannot be extracted from a digital file beyond what was originally captured.

For pages containing modern printed text only, high quality scans produced according to well-understood standards are likely to remain sufficient for future needs; unless future scholars develop a deep interest in the grain of journal paper or other microscopic features, there is little meaningful additional information to be captured from these pages. Therefore, the need to retain text-only journals in print for such purposes is minimal, such as the grayscale-driven OCR mentioned above.

The digitization of images has proven to be more problematic to conduct at scale. While standards have improved markedly, even today subject experts occasionally report that image quality is inadequate to replace print. For this reason, journals containing figures and images of significant importance to scholarship may have greater need of being eventually redigitized. This indicates that there may be a greater need to retain image-heavy journals for potential redigitization or, if re-scanning is prohibitively expensive, for somewhat regular, if infrequent in many fields, scholarly use. [20]

Digital preservation

Observers often consider the retention of print versions of scholarly materials as a preservation backup, enabling redigitization in case of some system failure of the digitized versions. To the extent this need exists, however, it should properly be seen as a failure of the digital preservation infrastructure.

Digital preservation solutions are designed to guard against the risks of system failure, evolution in file formats, and changes in ownership and governance. To guard against these risks, digitized backfiles should be deposited in a community preservation solution.[21] Indeed some preservation initiatives already include a substantial amount of digitized backfiles in addition to born-digital materials.[22] Digitized materials preserved in this way are extremely unlikely to require redigitization from source materials due to data loss or other failures.[23] Further, CRL's auditing efforts are providing objective assurance of the quality of such preservation solutions.[24] For digitized journals deposited in an appropriate digital preservation solution, there is essentially no additional value associated with retaining print as a preservation backup against system failure.

Unfortunately, not all digitized materials are deposited with digital preservation solutions. Insufficiently preserved digitized materials are far more fragile and subject to loss. In addition, the digital preservation processes with which digital resources are managed may not always be clear, leaving the library

[20] There is a fairly extensive literature examining images in online journal collections, although anecdotal evidence suggests that many of these concerns have abated in recent years for scientific material if not always for art images. Xiaotian Chen, "Figures and Tables Omitted from Online Periodical Articles: A Comparison of Vendors and Information Missing from Full-Text Databases," *Internet Reference Services Quarterly* 10, no. 2 (July 2005): 75-88; Jacquelyn Marie Erdman, "Image quality in electronic journals: A case study of Elsevier geology titles," *Library Collections, Acquisitions, and Technical Services* 30, no. 3-4 (December 2006): 169-178; Carolyn Henebry, Ellen Safley, and Sarah E. George, "Before You Cancel the Paper, Beware—All Electronic Journals in 2001 Are Not Created Equal," *The Serials Librarian* 42, no. 3 & 4 (July 2002): 267-273; Lura E. Joseph, "Image and figure quality: A study of Elsevier's Earth and Planetary Sciences electronic journal back file package," *Library Collections, Acquisitions, and Technical Services* 30, no. 3-4 (December 2006): 162-168

[21] Best practices are not uniformly settled on which files should be deposited. Some believe master image files from scanning are critical, while others believe that delivery-quality files can suffice..

[22] For example, at the time of this writing, over 4 million of the 13.1 million journal articles deposited in Portico were retrospectively digitized, a figure that is expected to grow with additional deposits.

[23] Although secure in most scenarios, some have expressed concern that a globally catastrophic event such as nuclear war or other electromagnetic pulse is likely to defeat these preservation efforts. While this may be the case, such events would have dramatic implications reaching far beyond the preservation of scholarly journals. As such apocalyptic events would have pervasive and massive impact on all aspects of modern society, planning for their eventuality should come from a national or international policy perspective, and doing so exclusively for scholarly journals makes little sense. For more information on this issue, see "Commission to Assess the Threat to the United States from Electromagnetic Pulse (EMP) Attack," http://www.empcommission.org/

[24] After conducting a series of reports on a variety of key digital repositories, CRL in 2009 is expected to complete detailed certification and assessment exercises for both Portico and HathiTrust. See http://www.crl.edu/content.asp?l1=13&l2=58&l3=181 for more information.

community unsure how much trust to place in a given digital collection. The CRL auditing process should therefore be applied widely and the results used to inform community decision-making.

When materials are not certifiably preserved in digital form, appropriately preserved print copies will be required. But ultimately, encouraging high quality digital preservation (and increased transparency about which digitized materials are preserved and how) is more broadly beneficial than backing up poor digital preservation practices by maintaining print in perpetuity.

Reliability of access

Beyond the preservation of the digitized files, however, it is also important to consider the circumstances associated with access. In the first place, the digital provider should be able to provide access with little or no unplanned downtime, so that users will have acceptable levels of access to the digitized version. Without this type of access reliability, withdrawing print versions locally might offer short-term benefits but long-term problems in providing access to users.

Terms of access are also important considerations. Can libraries trust that pricing and other terms of the license will remain reasonable in the future, either directly through the terms offered by the provider or through some form of post-cancellation access? For those digitized collections where such trust cannot be generated, the library system should retain some versions of print for potential competitive digitization, as a hedge against monopoly behavior by the digital provider.

Scholarly needs

In addition to the need to retain print versions as a base from which digital copies are produced, there may be some needs that, at least given current scanning and access mechanisms, require access to the print original. Such needs are understood to be almost completely restricted to images rather than text, given the issues with standards for image scanning discussed above. Little good information exists on the prevalence of scholarship that requires access to print images, and more research on this topic could clearly be beneficial.

In addition, idiosyncratic scholars may come to have interest in arcane features of text-based materials that are insufficiently captured by digitization. Such needs are edge cases and are correspondingly difficult to predict. Certainly, some rare book libraries or special collections will want to accession exemplars, at least, of what have been common scholarly journals, to enable book historians and other scholars to have access to this form of publishing.

While it is impossible to predict future scholarly needs with certitude, overall data from the University of California's print repository of JSTOR-digitized journals indicate that scholarly uses of this set of print originals are rare indeed.

Campus politics

Finally, campus politics may complicate preservation decisions. For many libraries, the risk that faculty will protest the removal of even the most rarely used print collections inhibits decision-making about print collections. As mentioned above, in some specific cases, faculty may have real needs for local access to print materials, and in these situations it may be wholly appropriate to maintain local collections until practices change. In most cases, however, faculty who exclusively rely on digital materials may still be wary about the system-wide elimination of print materials. Faculty members are significantly more likely to strongly feel print must be preserved in libraries somewhere than they are to feel it must be

preserved on their own campus.[25] The ability to point to the guaranteed availability of (and a mechanism for access to) print somewhere in the system would help the library to assuage the concerns of these key stakeholders without limiting local decision-making. Due to a decline in faculty interest in print preservation both locally and remotely in recent years, the political necessity of maintaining even remote access to print collections will probably remain a requirement only in the medium term.

COMMUNITY PRESERVATION REQUIREMENTS

We developed the foregoing rationales in order to estimate the value of community attention to print preservation generally and to analyze its specific characteristics. These rationales indicate a variety of community requirements for a print preservation system, such as the conditions that would require certain types of materials to be kept for longer or shorter periods of time. By determining what should be preserved, under what conditions, and for how long, we can ultimately determine where preservation is currently adequate and where it might need to be further strengthened.

An important component of the development of a community preservation strategy is the definition of time periods for which materials must be kept. It must be emphasized that the time periods we discuss here are lower bounds, representing a minimum amount of time for which at least one copy of an item must be reliably preserved in order to ensure that community needs can be effectively met. This does not indicate that after the time horizon has passed, remaining copies should or will be deaccessioned; rather, at that point, the ongoing needs for print materials should be reassessed. Our suggestions are aimed exclusively at ensuring that print versions remain available for *at least* as long as they are concretely needed. In any case, that the community gain some consensus on time horizons is far more important than the estimates that we propose here.

Scenarios

What to retain, and for how long, are highly interrelated. Even a simple set of characteristics to define a similar set of materials introduces complexity. For example, assume that materials have been digitized at a high rate of quality, are being actively corrected as errors are reported, are preserved digitally according to community best practices, are text-only, and are provided under reliable terms. This assumption integrates key elements of the rationales from the previous section into a scenario, labeled the "Ideal Scenario" in Table 1. Under these conditions, the rationales for print preservation presented in the previous section indicate a clear need for the continued availability of at least a limited set of the print originals, especially because of the need to have access to print versions for error correction. For the JSTOR-digitized journals it seems that error reporting is significant at first, and then begins to fall off after five years of online availability. Based on the rationales, we suggest the need for the community to have access to one or a small number of copies of the print version for at least 20 years. Beyond this point, however, the benefits of investing heavily in the preservation of this sort of print materials are less clear and seem likely to exceed the costs of doing so.

[25] In Ithaka's 2006 study, 57.6% of faculty indicated that they felt strongly that *some* libraries should maintain print copies of materials, while only 41.1% indicated that they felt that *their* library should do so. Housewright and Schonfeld, *Ithaka's 2006 Studies of Key Stakeholders in the Digital Transformation in Higher Education*

Table 1 The Characteristics of Four Exemplary Scenarios - and Their Implications

Key characteristics and their implications	Ideal Scenario	An Inadequate Digital Preservation Scenario	An Image Intensive Scenario	An Inadequate Digitization Quality Scenario
Digitized with high standards of quality?	Yes	Yes	Yes	**No**
Errors are actively being corrected?	Yes	Yes	Yes	**Yes**
Digital copies are reliably preserved?	Yes	No	Yes	**Yes**
Image-intensive?	No	No	Yes	**No**
Terms of provision are reliable?	Yes	Yes	Yes	**Yes**
How should individual libraries handle these materials?	These materials are suitable for local withdrawal at all types of libraries.	Without adequate assurance of preservation and accessibility, reliance on digitized versions is impossible for most libraries	Images may be needed for access purposes and present an unknown for preservation – Strongly consider retaining locally	**These materials may be suitable for withdrawal at some libraries, but perhaps not at a research library.**
Time horizon for ensuring that some print copies are retained across the community?	20 years, during which libraries should collaborate in upgrading the quality of the digitized versions.	n/a	n/a	**100 years, during which time materials would be re-digitized**
Level of assurance that at least one copy remains after the stated time horizon?	High – 99%+	n/a	n/a	Very high – 99.99%+

For some combinations of characteristics, our analysis cannot provide system-wide assurance that print withdrawals are acceptable. For example, in an inadequate digital preservation scenario, it is impossible to rely on the digitized version for access over the long-term. Consequently, no library that requires access to a certain set of content can responsibly rely on the digitized version alone. Adequate print must remain available to provide for access in case it is needed at some future point.

Similarly, in an image-intensive scenario, image-reliant scholars are likely to require access to the print versions of some journal sets that are image-intensive, and some such journal sets will experience a greater potential need of being redigitized. Determining the time horizon for these materials will be more speculative, because we do not have good estimates for when digitization costs for images are likely to allow for the highest fidelity digitization. Many libraries whose users require immediate access to image-laden journals may find the digitized version alone to be inadequate. Sufficient print must remain available to provide for access in case it is needed at some future point, and again the data available does not permit our analysis to provide system-wide assurance that print withdrawals are acceptable.

But even if digitized versions can be relied upon for the long term, there are some scenarios in which a greater level of print preservation of primarily textual materials is called for. For example, in an inadequate digitization quality scenario, we must assume a longer period of error correction, perhaps even

a comprehensive redigitization at some future point. And, given the certainty that significant redigitization will need to take place, we believe that it is correspondingly important to increase our level of assurance that print copies remain available. We therefore suggest a 100 year time horizon (which is surely conservative, but appropriately so given the poor quality of digitization) and an assurance threshold of 99.99%. Note that we are not quantifying the difference between acceptable digitization quality and inadequate digitization quality – one of our recommendations is that the preservation community develop formal standards and thresholds that allow for assessing the quality of digitization.

A key challenge for the community is to distinguish transparently and explicitly between those materials for which a managed drawdown might be reasonable and those for which widespread overlap remains important. In this section, we have suggested two exemplar scenarios in which some form of managed drawdown might make sense, and two in which it does not, based on the information to which we currently have access.

Where consensus can be achieved on time requirements, they should be seen as minimum requirements rather than withdrawal targets, and the system should incorporate a mechanism to re-evaluate them and manage either a further drawdown or a recommitment to retention at some future point. These time horizons and the assumptions that underlie them can certainly be debated, and we are eager to engage with others' ideas of the appropriate amounts of time for which print materials must be retained in order to accomplish specific community preservation goals. But the presence of time horizons enables libraries to make finite and more clearly defined commitments to print preservation, which can be renewed, passed on to other members of the community, or given up as appropriate when a milestone is reached.

Quality and availability

Beyond time requirements, it is also important to consider under what conditions journals will be maintained and access provided. Journal collections, even those that do not circulate, often are missing articles, issues, or even volumes. Providing quality assurance for these collections would be especially important if a large-scale redigitization is expected at some point in the future, so that complete backfile sets can be readily accessed. On the other hand, it might prove more cost-effective to retain a greater number of copies without undertaking quality assurance procedures, providing some statistical assurance that enough copies remain across collections to feel confident that complete runs can be assembled.

Moreover, as the number of print copies is reduced across the system, individual copies take on greater value through their increased rarity. Materials may be held explicitly for preservation under restricted usage paradigms, and ownership of print materials may become much more concentrated. It is therefore important to consider how such materials may be accessed for both scholarly and digitization-related needs. Some restrictions on physical access may certainly be warranted, out of concern for the long-term survivability of the physical artifact. For example, to obtain access scholars may be required to demonstrate a compelling need for physical access such that a digital surrogate is insufficient for their purposes. Digitizers might be required to utilize non-destructive scanning techniques that will not compromise the print original. Restrictions must therefore be crafted with the recognition that a print collection only has value if it can be used when necessary. These restrictions should enable access for scholars and digitizers as needed under reasonable and non-discriminatory terms, rather than limiting accessibility to only those affiliated with certain institutions or entities or blocking certain types of uses entirely.

MODELING PRINT PRESERVATION

As the foregoing sections have made clear, requirements for print preservation will vary based on a number of factors. Among the key factors would is the time horizon for print preservation and the degree

of assurance required. An established and well-defined system for print preservation would address community preservation requirements and allow for those materials incorporated into it to be withdrawn by other libraries. Ithaka S+R commissioned Candace Yano, a professor of industrial engineering and operations research and in the Haas School of Business at UC Berkeley, to perform a research study aimed at determining how such a system might work.[26]

Yano and her team examined the requirements to ensure the availability of at least one complete copy of a journal backfile set over a certain time horizon. The implicit assumption is that regular access needs will be met through a digital surrogate. Because little good data are available, example calculations assume an annual rate of "loss" of 0.1% for non-circulating "dark archived" collections, along with riskier circulating collections whose lose rate is estimated at 0.5%.[27] The underlying model can accommodate any assumptions on all these variables, and the Yano team developed a simple tool to allow for evaluations as needed.[28]

To provide a simple example, assume that the print journals will be validated at the page level, ensuring the perfect quality of the archive at the first point in time. Also, assume that the journals will not circulate, meaning that many sources of damage can be effectively eliminated. Under such conditions, the number of print copies required under two of the scenarios discussed in the previous section are illustrated in Table 2. If, as we have suggested, requirements for preservation vary depending on certain types of material types, so too does the number of print copies required to provide acceptable levels of preservation vary.

Table 2 The Yano Team's Simple Model Applied to Two Scenarios

Scenario	Time Horizon	Probability of Success	Number of "Perfect," Uncirculating Copies Required
Ideal Scenario	20	99%	2
Inadequate Digitization Quality	100	99.99%	4

As Table 2 depicts, two dark, page-verified copies are sufficient in the Ideal Scenario to offer 99% confidence that at least one copy will survive for the twenty year time period. But other scenarios – such as described above – may require a larger number of such copies. Recognizing that assembling page-verified collections is extremely costly, and the assembly of a significant number of page-verified copies may be simply infeasible, the Yano team also developed by which similar levels of reassurance can be provided with a combination of page-validated dark copies, as described above, volume-validated circulating copies. If dark archive copies are damaged or destroyed, open collection copies can be used to replace or reconstitute them. At least two dark, page-verified copies are required so that if one is lost or

[26] Candace Arai Yano, Z.J. Max Shen, and Stephen Chan, "JSTOR Seeks Efficiency and Security for Print Backups of Online Journals," *Interfaces* (2009, under review).

[27] This figure is roughly the loss rate paid by insurance companies for library collections, including both circulating and non-circulating material, after deductibles. A more accurate loss rate might be derived from this statistic by adding back in deductibles, account for the difference between circulating and non-circulating copies, and account for losses that are never identified by libraries.

[28] Yano's spreadsheet-based tools are available at http://ieor.berkeley.edu/~shen/SurvivalPerfectCopies.xls for the "simple model" presented in this section and at http://ieor.berkeley.edu/~shen/HybridSystemAnalysis.xls for the hybrid model discussed in the following section.

damaged, one will remain against which the circulating copies may be verified, so we cannot apply this model to simplify the Ideal Scenario. But in the case of an inadequate digitization quality scenario, such a hybrid approach may be valuable. Figure 1 plots the number of dark and circulating copies needed for such a scenario, demonstrating how a varying number of dark archive copies may be adequately supplemented with circulating copies.

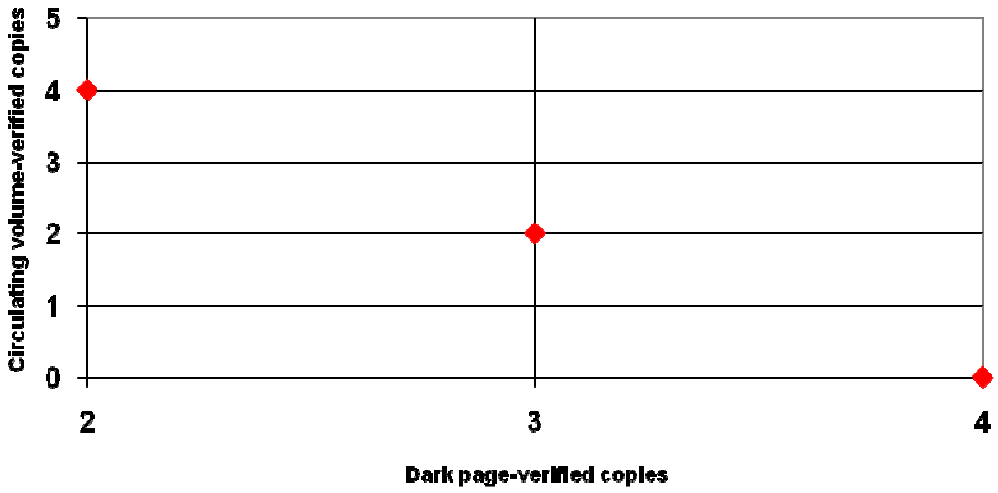

Figure 1 Copies Required for an Acceptable Preservation System for the Inadequate Digitization Quality Scenario

For such a hybrid model to be put into practice, a minimum of two dark, page-verified copies must still be assembled. But the costs associated with the development of additional such copies can be obviated if sufficient circulating copies, verified at the volume level, are also present. These circulating copies, however, must be held as a part of a preservation system; clear commitments must be made to maintain these collections for the target time horizon. As described above, few if any libraries have transparently described their preservation commitments for various parts of their collections, and there is currently no way to effectively determine which materials are likely to be widely maintained and which may be at greater risk. To enable a hybrid preservation model, protocols must be developed to enable the simple sharing of information about what circulating collections can be considered as a part of this preservation system. If several large research libraries were to use such a protocol to formalize their commitments to maintain certain parts of their collections, it is likely that a reasonable number of circulating collections could be easily assembled to contribute to a hybrid preservation model.

The Yano team's modeling simplifies print preservation, offering two options for preservation conditions, when in reality materials can be in many different types of conditions and preservation circumstances vary in numerous ways. Additional modeling in this vein, perhaps complicating the Yano work to allow for additional conditions, might be undertaken. In particular, a model that does not rely on page-verification would be of great value, as this essential step in Yano's model poses potential funding challenges. Still, for the first time a scientific approach to print preservation has been developed, allowing the community to make appropriate collection management choices. While the Yano team's modeling may not cover all the conditions under which print can be appropriately withdrawn from local collections, it allows us to conclude with confidence that certain text journals' print versions can be withdrawn immediately – and responsibly – from library shelves.

JOURNALS WITH IMMEDIATE WITHDRAWAL POTENTIAL

A large and rapidly increasing number of journals are nearly universally accessed in digital form. For some of those materials reflective of the Ideal Scenario we have been discussing, taking a new approach to collections management may be in order today or in the near future. Currently, most print copies of the journal materials that meet the terms of the Ideal Scenario are held in a completely uncoordinated way across the library system, without any page-verified copies, and the development of a system that effectively organizes system-wide print holdings is necessary. The JSTOR-digitized journals offer a unique opportunity because many of its journals fall into the Ideal Scenario.

JSTOR has been a leader in digital preservation, uses high-quality digitization practices, corrects digitization errors as soon as they are discovered, and is seen as a reliable digital provider. It was created to allow libraries to rely on it for access and preservation purposes and thereby to free up shelf space, and it has become trusted by the community for its standards and sustainability.[29]

Consequently, the JSTOR-digitized text-only journals seem to offer an ideal opportunity to consider a widespread draw-down of print collections across the library system. The Ideal Scenario calls for a minimum of two page-verified copies, and JSTOR has commissioned the creation of two page-verified dark archives of its holdings, one at Harvard and one at the University of California. Hundreds of titles have been preserved in both repositories as of 2009, but efforts to populate them lag the online release of the digitized versions to some degree.

Other efforts have attempted to preserve the JSTOR-digitized journals. The Five Colleges in Massachusetts preserves a set of the print versions on a shared basis to allow individual libraries to provide print access after withdrawing their locally held copies. Other shared print facilities have undertaken similar efforts, for similar purposes. CRL's internal print archive project, along with its distributed archive initiative, are experimenting with other types of mechanisms for preserving the same print materials under other types of arrangements. If there were an efficient protocol to share information about the completeness of these efforts, the conditions of the materials they hold and the conditions under which they hold them, and their long-term stability and sustainability, [30] the level of assurance that they provide could be added to the level of assurance provided by the JSTOR-commissioned print copies. Although according to Yano's model, these copies may be in excess of what is strictly necessary, they provide an even higher statistical likelihood of having at least one copy remaining and for a longer time horizon. Perhaps more importantly, they may address whatever ongoing access needs remain.

In sum, we believe that the text-only JSTOR-digitized journals are becoming subject to sufficient print preservation that local copies are no longer required, even at the research libraries whose retention and preservation commitments would traditionally have served as key components of the preservation framework.

Still, we do not mean to suggest that any given library should take the decision to withdraw. Such a choice to withdraw presents political conditions that vary from institution to institution, and we hear again and again of campuses that have provoked sharp faculty or student reactions to their decision to withdraw print backfiles.[31] Our analysis suggests that there are some print journals that are essentially risk-free from

[29] Roger Schonfeld, *JSTOR: A History* (Princeton, New Jersey: Princeton University Press, 2003).

[30] Stability is very important, as we will discuss below, with at least one print repository of JSTOR materials having already been disbanded. See http://www.crl.edu/content.asp?l1=13&l2=19&l3=35&l4=63.

[31] For example, the decision to deaccession JSTOR journals at Cal Poly Pomona was very controversial among both librarians and faculty. See Scott Carlson, "Library Renovation Leads to Soul Searching at Cal Poly," *The Chronicle of Higher Education*, September 1, 2006, http://chronicle.com/free/v53/i02/02a05901.htm

a preservation perspective and therefore should be considered for withdrawal before other backfiles are similarly considered. This may prove reassuring to campus stakeholders in explaining the collections management choices a library is pursuing and setting them in system-wide context.

While this section has focused on the JSTOR-digitized text journals, we do not mean to imply that other print journals do not meet the thresholds described to enable immediate withdrawal. We hope that the specifics of this scenario will help librarians or digital providers to conduct a similar analysis that indicates adequate reassurance exists for other collections as well.

INCREASING THE WITHDRAWAL POTENTIAL OF OTHER JOURNALS

Indeed, given the variety of print preservation initiatives in place or under development, it is quite possible that other journals are acceptably preserved in print format relative to community requirements. However, it is impossible to determine readily that such preservation exists, because protocols for exchanging such information have not been developed. Moreover, many journals should not be subject to withdrawal for more fundamental reasons. In this section, we examine the journals that may not be adequately preserved relative to community requirements, with the objective of providing recommendations for improved preservation that will allow for eventual withdrawals.

Digitization quality

As we have seen in previous sections, to design a systemwide collection management strategy for print collections requires that we know something about how they were digitized and in what digital preservation systems they are deposited. Many organizations undertaking digitization are vague or even secretive about the processes they use and the quality that is achieved. Consequently, it is impossible to design an efficient print collections management strategy. In the first instance, therefore, organizations undertaking digitization should release far more information about their practices and quality standards, preferably in a fairly standard fashion allowing for effective comparison.

Regardless of the quality that any given process might expect to achieve, errors are inevitable in large-scale digitization processes. Quality control can also be implemented after the fact, using user reports to identify errors and initiate a repair process. We earlier reported JSTOR's practice of repairing such errors after they are reported. While error-reporting to JSTOR is initiated through a multi-purpose feedback form, Google Books offers a Feedback tool that is solely designed to collect error reports, although it is not clear how if at all Google intends to use the data gathered through it to increase the quality of its offerings.

Gathering these data is only the first step. One can imagine a mass digitizer, for example, developing partnerships with libraries that would manually scan missing or otherwise problematic pages from a list that it provided them. With appropriate tracking of these processes and their outcomes, the quality of a project that initially pursued mass-digitization solely for access purposes might eventually be brought up to preservation quality. As a result of such processes, libraries would be able to withdraw a greater share of collections, thereby providing an important incentive for their participation.

Reliability of the provider

In order for access to shift to the digitized version not only for the short-term, but also for the long-term, libraries must ensure the reliability of the provider. Reliability can be determined through a number of means, including governance structure or terms of the license agreement. License terms that ensure that access will remain available post-cancellation, as well as provisions for efficiently realizing such access,

are becoming an increasingly important mechanism for ensuring the long-term reliability of content licensed from traditional publishers.

Digital preservation
All digitized journal backfiles should be deposited into digital preservation solutions. As scanning errors are repaired, systems for appropriately ingesting updated files will also be important. Providers of digitized journal backfiles should transparently provide greater detail about how these materials are digitally preserved, including the results of CRL audits.

Images
A significant share of journal content contains charts, graphs, data tables, photographs, drawings, models, and other types of images, with importance across numerous disciplines. Too little is known at this point about the share of the intellectual content of an image that is captured by existing scanning and format standards. In addition, given that color images likely have some variance across a print run due to the nature of the printing process, it is not clear that there even is a single "perfect" source that a digitized version should attempt to replicate. More research is needed in these areas in order to determine whether image digitization can hope to provide a suitable preservation-quality substitute for print and if so for what image types and under what conditions.

Page verification
Regardless of the quality of the quality of the digitization, the Yano models require at least two page-verified copies of the print in order to provide adequate reassurance to allow widespread withdrawal of print. The most promising direction forward would be to identify uncirculated complete copies from publishers that could be spot-checked in some way. Otherwise, it would probably be necessary to build a page verification system up from existing groups of libraries, such as systems and consortia and especially existing shared print repositories.

The preservation of adequate copies of all backfiles in dark archives would be an enormously expensive and challenging undertaking, because assembling, validating, and storing these materials would require monumental investment. It is challenging to imagine how such a model could be effectively funded given the decentralized nature of the library system. The JSTOR –University of California partnership, which has page-verified one of the required copies for a collection of about 28 million pages, has cost $1.7 million. This is relatively low when allocated across thousands of libraries participating in JSTOR, but it is not clear if other central entities might be willing to invest in such print journal preservation, or if libraries would be willing to band together to fund such investment. Assembling the resources to undertake the page verification role and apportioning the responsibility for doing so fairly is a challenge that faces the community. Questions remain about whether publishers or libraries would contribute to the costs of such an effort, and how some form of sustainability would be assured. And, there is little evidence to suggest that the community could organize the resources needed without the involvement of a central catalyst like JSTOR, and while organizations from OCLC to CRL to HathiTrust could conceivably serve in such a role, none has yet taken up the mantle.

BUILDING A SYSTEM

In earlier generations, two factors made it impossible for libraries to move away from local holdings of print versions: the technology for information-sharing was immature and the organizational structures for collaboration were inadequate. More recently, the technology for information-sharing has matured substantially, and even if the infrastructure for information-sharing is imperfect, it could be readily developed. The larger problem remains organizational: how can inter-institutional collaborations

appropriately apportion responsibility, and provide revenues, to meet the need for some number of print copies to be retained for at least limited periods of time?

One of the most significant conceptual problems facing libraries that would like to band together to preserve print collections is the nature of their decision-making process. Many libraries are most likely to collaborate on print management issues only when faced with an acute local space shortage, and consequently some have been tempted to withdraw from collaborative initiatives if that space shortage is resolved via expansion. A new strategic approach to local print collections, recognizing their dramatically changed role both in meeting user needs and in contributing to preservation, must be adopted by a library in order to serve as an effective long-term partner in collaborative enterprises.

A strategic approach would allow collaboration to take any of a number of different directions. Many observers in recent years have looked to print repositories, which are potentially a promising approach so long as organizational design is carefully considered. The history of print repositories provides some useful indications about how they might be organized to serve the systemwide objectives discussed in this report. The first generation of print repositories in the 1950s included the New England Deposit Library (NEDL) and the Midwest Interlibrary Center (MILC). Neither of these print repositories served for very long as the main collections storage center for its participating libraries. For these initiatives, long-term sustainability for their originally intended purposes would have required the design of more effective organizational models coupled with incentives that were adequate to motivate ongoing participation.[32]

But there are other models as well, such as CRL's print preservation efforts. Internally, CRL has worked to assemble a copy of the collections based on the drawdowns from a variety of research libraries, serving in this respect essentially as a print repository. But in addition, it has developed a program whereby several of its members can formalize their commitment to retain print versions in their campus environments. The lessons being learned from this undertaking will be important in helping the community understand whether distributed collections can be adequately secured or whether a stronger organizational structure is required.[33]

Another model to look towards is the UK Research Reserve, piloted in 2007-08 and launched in 2009.[34] The UKRR is a coordinated approach across the higher education system and in partnership with the British Library funded by the Higher Education Funding Council for England to ensure expanded access to and secure preservation of print materials across the community. It is a major positive development, providing for less than the levels of reassurance imagined by Yano's findings but more than would be available otherwise, and it could be seen as a component of a broader system, since preservation is a challenge that cuts across national boundaries.[35] Other countries, such as the US, will be less likely to develop the central coordination and funding that has benefitted the UKRR.

In the longer term, binding together individual repositories and library commitments for wide varieties of different types of materials will prove to be a key challenge. How will responsibility be apportioned and commitments vocalized? Appropriate agreements across repositories are needed to ensure their long-term

[32] Rachel Burstein and Roger Schonfeld, "The First Generation of Print Repositories," *Libraries and the Cultural Record* (2010) forthcoming.

[33] See http://www.crl.edu/content.asp?l1=13&l2=19&l3=35 for more information.

[34] A wealth of very useful information can be found at http://www.ukrr.ac.uk/.

[35] There is interest in these issues in other countries such as Australia. See for example Steve O'Connor and Cathie Jilovsky, "Approaches to the storage of low use and last copy research materials," *Library Collections, Acquisitions, and Technical Services* 32, no. 3-4 (2008): 121-126

commitments.[36] In addition, how will participants ensure that appropriate numbers of items are preserved and made accessible as necessary over time? Lightweight systems and protocols are needed to share information about retention commitments, page verification, and lacunae in order to enable appropriate decision-making.[37] Print repositories along with major research libraries should convene to build out such coordination among themselves as soon as practicable. Modest staffing may be necessary to coordinate their work.

Finally, even in the absence of a fully formalized system, sharing greater information about local retention commitments would have significant value by allowing other libraries to have a better sense of relative levels of risk and assurance. Thus, information-sharing – which is possible today – should be pursued even if the community concludes that the organizational structures for strong collaboration cannot be created.

RECOMMENDED ACTION STEPS

We break out recommendations into three sets of action steps:

Action steps for journals with immediate withdrawal potential

ONE: *Libraries can responsibly withdraw the print versions of certain journals.* Our analysis suggests that libraries can withdraw certain print versions, such as those JSTOR-digitized text-only journals that are held in two print repositories, without in any way posing risks to community preservation requirements.

Action steps to increase the withdrawal potential of other journals

TWO: A standard protocol should be developed for the public expression of print retention commitments. These commitments should have several features:

- First, the library or repository should explain precisely which materials will be retained.
- Second, the library or repository should acknowledge the present condition of these materials and fill in any volume-level gaps.
- Third, the library or repository should state the conditions under which it will maintain these materials as well as rules for usage.
- Fourth, the library or repository should state the time period for its commitment and when and how it will reevaluate this commitment.
- Fifth, the library or repository should incorporate plans for how to redistribute these materials to other interested parties should they no longer be willing or able to maintain the commitment in the future.

THREE: The holdings of existing print repository efforts should be analyzed collectively. Based on the information provided via the standard protocol, an analysis should be conducted of how many volume-verified copies currently exist for any given journal volume. This exercise would help the community take stock of what additional steps might be required.

FOUR: *A study should examine the level of assurance provided in the absence of page-verification.* Adequate data do not yet exist to quantify the level of assurance of a preservation system that lacks page-verification. Given the challenges to expanding page-verification, it may be hoped that adequate

[36] Further discussion of this theme can be found in Constance Malpas, "RLG Partnership Shared Print Collections Working Group: Shared Print Policy Review Report" (OCLC Research, January 2009), http://www.oclc.org/programs/publications/reports/2009-03.pdf

[37] For example, OCLC prototyped the Cooperative Collection Management Trust to provide information on retention commitments, and it could analyze commitments relative to a target number of copies in an eventual system.

assurance can be provided at the volume level without too many copies being required. But, if page-verification remains requisite, the community will have to identify models for page-verification, or modify its preservation requirements.

FIVE: Several large research libraries should commit to retain print collections. If any large research libraries were positioned to commit to retain collections that in any case they would be highly unlikely to withdraw (using the protocols from Recommendation TWO), these commitments could almost immediately address any needs for circulating print copies as a part of a hybrid solution. An aggressive effort to bulk up CRL's distributed print archive is one possible direction, though other approaches might also be considered.

SIX: A study should examine the extent to which there will be scholarly dependence on print versions of digitized materials. While studies have already found that there is very little bona fide scholarly need for access to print versions of most digitized materials, it is believed that image-intensive materials may constitute an important exception. Is this the key dividing line or are there others? Would a higher quality digitization process for image-intensive materials be unreasonably expensive relative to the marginal benefits? Are there other approaches that might empower image-intensive fields to make a more complete transition to the electronic environment?

SEVEN: Digitization should follow accepted standards, digitized materials should be deposited into sustainable digital preservation solutions, and publishers and vendors should provide greater transparency about such standards and practices. This will help the community to categorize the quality of and risk to collections, which is a prerequisite to gaining consensus around appropriate preservation approaches for each category. In addition, based on awareness of actual practices, the library community should encourage digitization programs wherever possible to utilize preservation-caliber scanning practices and to deposit the resulting digital files (including master image files) in preservation archives according to community standards. If digital editions are inadequate for long-term access to materials, then print journals take on far greater importance and correspondingly more complex print preservation systems may be required.

Building a system

EIGHT: The library community should convene around broader questions of print preservation. Will other organizations beyond JSTOR serve as the catalyst for preserving print titles of interest or will the burden fall to individual libraries or groups of them? What rationales and system requirements will be pursued? Will page verification be necessary or will another approach be pursued? Existing consortia and repositories must engage with the preservation-minded research libraries and determine how to gain confidence that a sufficient number of print copies of relevant materials will be preserved against community needs. The appropriate solution may involve some combination of centralized archives and coordinated distributed copies, but the current overlap-driven model cannot be relied on to adequately guarantee that community needs for print materials will be met.

NINE: *Lightweight yet reliable mechanisms are needed for the automated exchange of information about preservation.* As Recommendation TWO indicates, individual libraries and community initiatives will increasingly need to make explicit preservation commitments for print materials. It is increasingly a challenge, however, to share these commitments in an automated fashion that enables effective decision-making. What level of detail must be shared and can standard, automated protocols be designed to share these data? While the requirement for such a system is clear for print preservation commitments, such a system will be most valuable if linked with data about digital preservation commitments as well.

TEN: Best practices should be formulated to help libraries make effective local collections management decisions in the context of community-wide actions and priorities. In contemplating the preservation of print collections, or their withdrawal, libraries need to partner effectively with faculty members and understand their needs and concerns. Disciplinary needs and sensitivities emerge as particularly important considerations for libraries.

CONCLUSION

For journal collections that are available digitally, the online version provides for virtually all access needs, leaving print versions to serve a preservation role and therefore be required in far fewer numbers. Based on our analysis of community needs for print original materials, we see at least a medium-term need to maintain some level of access to print originals in the library community, although our analysis suggests in certain scenarios that long-term or perpetual preservation of many print materials may not be necessary. Consequently, certain print journals can be responsibly withdrawn today, and with concerted effort it should be possible to steadily increase the journals subject to withdrawal. On the other hand, we have also reviewed a number of situations, such as image-intensive titles and digitized version that are subject to inadequate digital preservation, in which the community will rely on print copies retained and preserved without system-wide coordination by hundreds of libraries that have long been the backbone of assured preservation and access.

Divining the right approach for any given title will not always be simple. Libraries that have historically retained collections over time face the real challenge of determining appropriate choices, not only for their local needs but also with an eye towards the system-wide considerations that this report has detailed.

We hope that collections managers and preservation librarians will feel inspired to work together on withdrawals. A managed drawdown of print that is no longer required will allow limited preservation resources to be redirected to higher priorities, not least of which are rare and unique collections.

In this report, we have not attempted to analyze the economic value associated with a shift to a system-wide print preservation system, nor have we attempted to compare such value against other priorities in an environment of constrained resources. Continuing the current decentralized approach remains an opportunity available to the community, acknowledging that an unmanaged drawdown will likely bring losses, just as happened with the newsprint transition to microfilm. We therefore hope to inspire healthy debate about not only the assumptions and analysis in the report but also about the prioritization of the print preservation problem more broadly. We are confident that by taking a pragmatic, well-reasoned approach to these issues, the community can achieve consensus around print preservation that allows for appropriate print collection management in the digital age.

WORKS CITED

Baker, Nicholson. *Double Fold*. New York, NY: Random House, 2001.

Bennett, Scott. *Report on the Conoco Project in German Literature and Geology*. Mountain View, CA: Research Libraries Group, 1987.

Burstein, Rachel, and Roger Schonfeld. "The First Generation of Print Repositories." *Libraries and the Cultural Record* (2010).

Carlson, Scott. "Library Renovation Leads to Soul Searching at Cal Poly." *The Chronicle of Higher Education*, September 1, 2006. http://chronicle.com/free/v53/i02/02a05901.htm.

Chen, Xiaotian. "Figures and Tables Omitted from Online Periodical Articles: A Comparison of Vendors and Information Missing from Full-Text Databases." *Internet Reference Services Quarterly* 10, no. 2 (July 2005): 75-88.

Cohen, Dan. "Google Fingers," June 26, 2006. http://www.dancohen.org/blog/posts/google_fingers.

"Commission to Assess the Threat to the United States from Electromagnetic Pulse (EMP) Attack." http://www.empcommission.org/.

Cox, Richard J. *Vandals in the Stacks? A Response to Nicholson Baker's Assault on Libraries*. Vol. 98. Contributions in Librarianship and Information Science. Westport, Connecticut: Greenwood Press, 2002.

Duguid, Paul. "Inheritance and loss? A brief survey of Google Books." *First Monday* 12, no. 8 (August 2007). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1972/1847.

Erdman, Jacquelyn Marie. "Image quality in electronic journals: A case study of Elsevier geology titles." *Library Collections, Acquisitions, and Technical Services* 30, no. 3-4 (December 2006): 169-178.

Gwinn, Nancy E., and Paul H. Mosher. "Coordinating Collection Development: The RLG Conspectus." *College and Research Libraries* 44, no. 2 (March 1983): 128-140.

Henebry, Carolyn, Ellen Safley, and Sarah E. George. "Before You Cancel the Paper, Beware—All Electronic Journals in 2001 Are Not Created Equal." *The Serials Librarian* 42, no. 3 & 4 (July 2002): 267-273.

Housewright, Ross, and Roger Schonfeld. *Ithaka's 2006 Studies of Key Stakeholders in the Digital Transformation in Higher Education*. Ithaka, August 18, 2008.

"Janus Conference on Research Library Collections, Cornell University, October 9-11,2005." http://www.library.cornell.edu/janusconference/index.html.

Joseph, Lura E. "Image and figure quality: A study of Elsevier's Earth and Planetary Sciences electronic journal back file package." *Library Collections, Acquisitions, and Technical Services* 30, no. 3-4 (December 2006): 162-168.

Macky, T. "Interlibrary Loan: An Acceptable Alternative to Purchase." *Wilson Library Bulletin* 63, no. 5 (January 1989): 54-56.

Malpas, Constance. "RLG Partnership Shared Print Collections Working Group: Shared Print Policy Review Report." OCLC Research, January 2009. http://www.oclc.org/programs/publications/reports/2009-03.pdf.

Mangan, Katherine S. "Packing Up the Books." *The Chronicle of Higher Education*, July 1, 2005.

Nichols, Stephen G., and Abby Smith. "The Evidence in Hand: Report of the Task Force on the Artifact in Library Collections." Council on Library and Information Resources, 2001. http://www.clir.org/pubs/reports/pub103/contents.html.

"North Dakota Last Copy Retention Policy." *The Unabashed Librarian*, no. 83 (1992): 24.

O'Connor, Steve, and Cathie Jilovsky. "Approaches to the storage of low use and last copy research materials." *Library Collections, Acquisitions, and Technical Services* 32, no. 3-4 (2008): 121-126.

Prabha, Chandra. "Shifting from Print to Electronic Journals in ARL University Libraries." *Serials Review* 33, no. 1 (March 2007): 4-13.

Schonfeld, Roger. "Commodity Collections: The Role of American Academic Libraries in the Maintenance of Knowledge, 1876-1900" presented at the Society for the History of Authorship, Reading, and Publishing, Halifax, Nova Scotia, Canada, June 2005.

———. *JSTOR: A History*. Princeton, New Jersey: Princeton University Press, 2003.

———. "The Role of Information-Sharing in Book Survivability in the United States, 1890-1940" presented at the Society for the History of Technology, Las Vegas, NV, October 2006.

Schonfeld, Roger C., Donald W. King, Ann Okerson, and Eileen Gifford Fenton. *The Nonsubscription Side of Periodicals: Changes in Library Operations and Costs between Print and Electronic Formats*. Washington, D.C.: Council on Library and Information Resources, June 2004.

Schonfeld, Roger C., and Brian F. Lavoie. "Books without Boundaries: A Brief Tour of the System-wide Print Book Collection." *Journal of Electronic Publishing* 9, no. 2 (Summer 2006).

Schottlaender, Brian E.C., Gary S. Lawrence, Cecily Johns, Clair Le Donne, and Laura Fosbender. "Collection Management Strategies in a Digital Environment." http://www.ucop.edu/cmi/finalreport/index.html.

Smith, Abby. "The Future of the Past: Preservation in American Research Libraries." Council on Library and Information Resources, April 1999.

The Commission on Preservation and Access. "Selection for Preservation of Research Library Materials," August 1989. http://www.clir.org/pubs/reports/lesk/select.html.

Weech, Terry L, Bryce L. Allen, and F. Wilfrid Lancaster. "The Last Copy Center Study. Illinois State Library Feasibility Study.." *Illinois Libraries*, no. 72 (1990): 44-50.

Yano, Candace Arai, Z.J. Max Shen, and Stephen Chan. "JSTOR Seeks Efficiency and Security for Print Backups of Online Journals." *Interfaces* (2009, under review).