A Model to Determine Optimal Numbers of Monograph Copies for Preservation in Shared Print Collections

Ian Bogus, Candace Arai Yano, Shannon Zachary, Jacob Nadal, Mary Miller, Helen N. Levenson, Fern Brody, and Sara Amato

In this study we developed a model and a spreadsheet tool for calculating, based on user input informed by available data, the probability of at least one usable copy of a monograph title surviving at various time horizons in shared print collections. The calculation incorporates four risk factors, which were assigned values based on research in the literature and our own studies. We applied the model to sample selected time horizons and risk tolerances, which suggests a minimum number of copies of a title needed for retention.

Introduction

Shared print library agreements offer a natural extension to research libraries' missions: they provide an extended pool of print resources to their user communities while attempting to secure the accessibility of each title well into the future. Many libraries are using networked retention commitments as part of their criteria when making collection management decisions. An evidence-based approach to determining retention targets— how many copies to keep, with the intent of limiting the probability of loss or irreparable damage to all copies of a title—has so far been lacking.

Retention commitments may serve a variety of goals. Fundamentally, they are intended to ensure access to a title through the term of the agreement. The duration of retention agreements differs between programs and ranges from as little as ten years to essentially unrestricted or permanent retention. In reality, it is not possible to guarantee that every title will survive in its

*Ian Bogus, Executive Director, Research Collections and Preservation Consortium (ReCAP), email: ibogus@princeton.edu; Candace Arai Yano, Distinguished Professor, IEOR Department and Haas School of Business, University of California - Berkeley, email: yano@ieor.berkeley.edu; Shannon Zachary, Head, Department of Preservation and Conservation, University of Michigan Library, email: szachary@umich.edu; Jacob Nadal, Director for Preservation, Library of Congress, email: jnad@loc.gov; Mary Miller, Director of Collection Management and Preservation, University of Minnesota Libraries - Twin Cities, email: memiller@umn.edu; Helen N. Levenson, Associate Professor, Collection Development Librarian, Oakland University, email: hlevenson@oakland.edu; Fern Brody, Associate University Librarian for Collections and Technical Services, University Library System, University of Pittsburgh, email: feb@pitt.edu; and Sara Amato, Data Librarian, Eastern Academic Scholars' Trust, email: sara@sharedprint.org. ©2023 Ian Bogus, Candace Arai Yano, Shannon Zachary, Jacob Nadal, Mary Miller, Helen N. Levenson, Fern Brody, and Sara Amato, Attribution-NonCommercial (https://creativecommons.org/licenses/by-nc/4.0/) CC BY-NC.

physical form. Humanity has lost texts throughout history, and it will continue to do so. In fact, it is likely that there are titles for which all copies now listed in WorldCat are unusable. While it should be rare, total loss will occur, and may occur in large enough numbers to cause discomfort. Shared print efforts cannot counteract every risk, but they provide a means to mitigate loss by providing better controls, distributing responsibilities and risks, and establishing intentional, multiparty oversight of collections management.

We look at large-scale issues as they affect consortia, geographic regions, and the implicit "collective collection" that libraries participate in through interlibrary loan (ILL) and similar consortial interlending systems. We focus on "average" books, for which the key retention decisions are how many and in which commonly available storage conditions to hold them.* This study is an investigation of retention decisions that we expect to be broadly applicable and that will ensure a high probability of survival of titles. The calculations are based on factors affecting every library and library collection, such as types of storage facility, prevailing or estimated risk of loss, or age and condition of subsets of collection material. One by-product of our study is to identify minimum viable levels of extant copies at which libraries have few or no options beyond retention of all copies and, implicitly, taking additional conservation and preservation actions to maintain those copies.

This is not a study of traditional preservation strategies, such as conservation treatment methods, protective enclosure designs, standards for preservation materials, or environmental controls. These preservation strategies will affect longevity and usability of specific groups of materials within the collection, and, therefore, the network-level retention targets of shared print networks. Although not the primary goal of this paper, the methodology that we present allows decision-makers to understand and quantify the impact of improved preservation strategies, at least in an approximate way, on enhancing the overall prospects of a print archive. As such, our methodology can aid in measuring the impact of preservation efforts, as well as determining appropriate resource investments and justifying them. We are suggesting a "Lots of Copies Keep Stuff Safe" strategy based on quantifiable metrics as part of an overall preservation strategy that is reliant on, and could affect the selection of, appropriate, traditional preservation strategies.

In this paper, we develop a quantitative model that enables us to identify tangible retention targets based on what is known about the key reasons that copies of book titles become lost, unusable, or otherwise unavailable over time.† We specifically include the following factors in the model: (1) on-shelf probability—the probability that an accurately-cataloged book is on the shelf or in a known location; (2) bibliographic record inaccuracy—the bibliographic record differs from the item known to be on the shelf; (3) annual loss rate—the annual rate at which copies are physically lost from the collection; (4) physical deterioration over time—the

* For our purposes, an "average book" is one that represents traits most commonly held in libraries. While in North America the average book may be in English and about 45 years old, one can define average books that exhibit specific traits, such as in Spanish, 100 years old, or having a specific construction. In our study we use a few different average books, mostly based on age.

† Throughout this study reference is made to titles and copies. In this context, a title is equivalent to the IFLA Functional Requirements for Bibliographic Records definition of a "manifestation," while a copy is the equivalent to the FRBR definition of an "item" (IFLA Study Group on the Functional Requirements for Bibliographic Records and International Federation of Library Associations and Institutions, eds., *Functional Requirements for Bibliographic Records: Final Report*, UBCIM Publications, new ser., v. 19 (München: K.G. Saur, 1998): 17–24). The loss of a copy means there are fewer copies of a given title available; the loss of a title means that no usable copies of that text have survived.

book is still available but degrades in usability over time, as a function of initial condition, use, storage conditions, and inherent vice. There are, of course, other reasons why copies of books are lost or irreparably damaged, but which we chose not to include in the model as discussed later in this paper.

The model and recommendations in this paper are intended to provide guidance as libraries and consortia are determining the number of copies of a title to retain to ensure with a high probability that at least one usable copy remains at the end of a specified time horizon; we refer to this probability as P1 for short. Although the model is designed to be general and flexible, and therefore can be utilized for other formats facing similar types of risk and physical degradation, the parameters chosen for our calculations are specific to print monographs.

When calculating the number of copies, it is assumed that said copies will have retention commitments. Commitments are necessary as, without them, copies may be withdrawn at any time, and they could not be relied upon to contribute to achieving the desired probability of at least one usable copy remaining.

Foundations for this Study

Our present work is founded on two articles. The first is "Optimising the Number of Copies and Storage Protocols for Print Preservation of Research Journal."[1] This paper by Yano et al. describes the first attempt at developing a model to aid in recommending retention of a given number of serial copies to ensure adequate preservation. The model incorporates some of the same factors that we consider in our study, but was designed for serials in particular, and for settings in which a few copies could be page-verified and placed in secure storage (e.g., off-site versus open library stacks), and backup copies could be committed and subsequently page-verified and moved to secure storage in the unlikely event that a copy in secure storage were lost or irreparably damaged. The need for page verification was motivated by JSTOR's need for clean (page-verified) copies to be scanned for inclusion in an electronic archive. In the setting we envision for this study, however, page verification for the monographs is not required. Similarly, for commonly occurring retention arrangements, a consortium would likely find it difficult to keep careful track of the number of monograph copies in secure, off-site storage and to coordinate replacement of a lost or irreparably damaged securely stored copy with one of the committed copies from elsewhere. Instead, for monographs, we envision that consortia would make decisions at a given time point regarding how many copies to keep in off-site storage and in stacks. Moving these copies at later dates could change the likelihood that one remains viable at the given horizon.

Furthermore, serials are structurally quite different from monographs and have historically been used differently. The authors of the Yano et al. study used a conservatively high annual loss rate of 0.5%, which means that, on the average, nearly 40% of a collection would be lost after a hundred years. This loss rate may be reasonably accurate for serials, where many works, in the form of articles, are bound together and shelved in open stacks. Serial volumes often show signs of excessive wear due to high use. Monographs do not appear to be used in the same way. Most monographs are used lightly if at all.[2] The ways that articles in a journal are used, defaced, and damaged differs from chapters in books. Because of the explicit consideration of page-verified copies in secure storage (with excellent environmental conditions), the Yano et al. study focused on the impact of loss of bound volumes and did not account for physical deterioration. Our context differs because the smaller usage rate of monographs means that

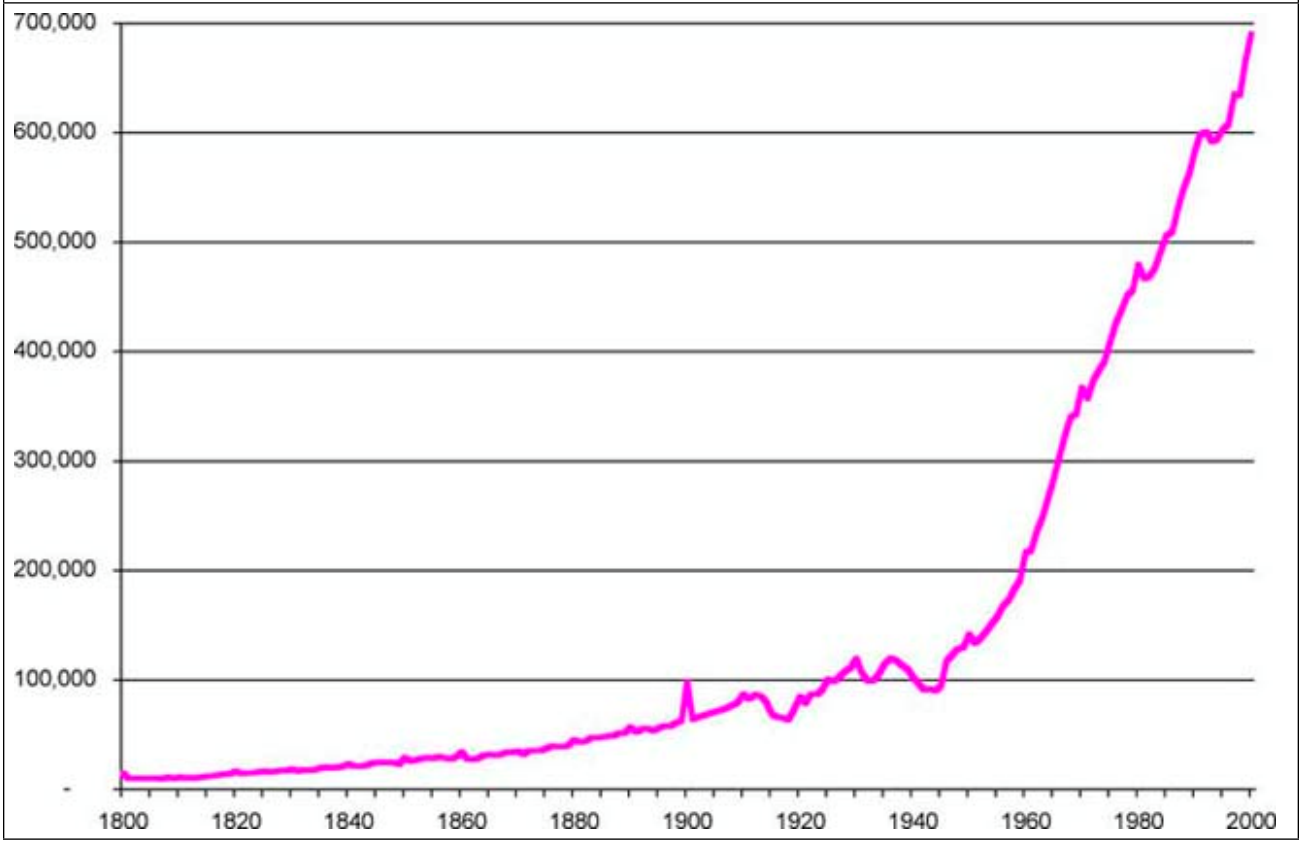
physical loss rates will be much lower. At the same time, the lack of both page verification and commitments to maintaining a minimum number of copies in secure storage, which typically also offers much better environmental conditions, means that physical degradation plays a larger role in the context of monographs. The model that we develop in this paper aims to account for these differences and other practical realities that apply to monographs.

This brings us to the second foundation article for this study: "Everything Not Saved Will be Lost," in which the authors attempt to identify the factors that will affect long-term retention in the context of shared print initiatives.[3] The holy grail is a well-reasoned recommendation on the number of retention commitments that each title needs. "The problem is that generating a recommended number is difficult, because to do so responsibly requires balancing several factors such as level of validation, condition, risk of loss, and long-term environmental storage, few of which have available data."[4] Solving this problem is possible by attempting to use rationally curated values and applying a mathematical approach.

Profile of Scholarly Print

For some, picturing a representative example of a book in a research library may conjure an image of an older, brittle item. The truth of the matter is that collections are, as a whole, much younger. There was a gradual increase in the number of titles published throughout the nineteenth century and a sharp increase after World War II. Schonfeld and Lavoie used

FIGURE 1
Print Manifestations by Year of Publication, 1800–2000
Reproduced by permission from Roger C. Schonfeld and Brian F. Lavoie, "Books without Boundaries: A Brief Tour of the System-Wide Print Book Collection," *The Journal of Electronic Publishing* 9, no. 2 (2006), https://doi.org/10.3998/3336451.0009.208.

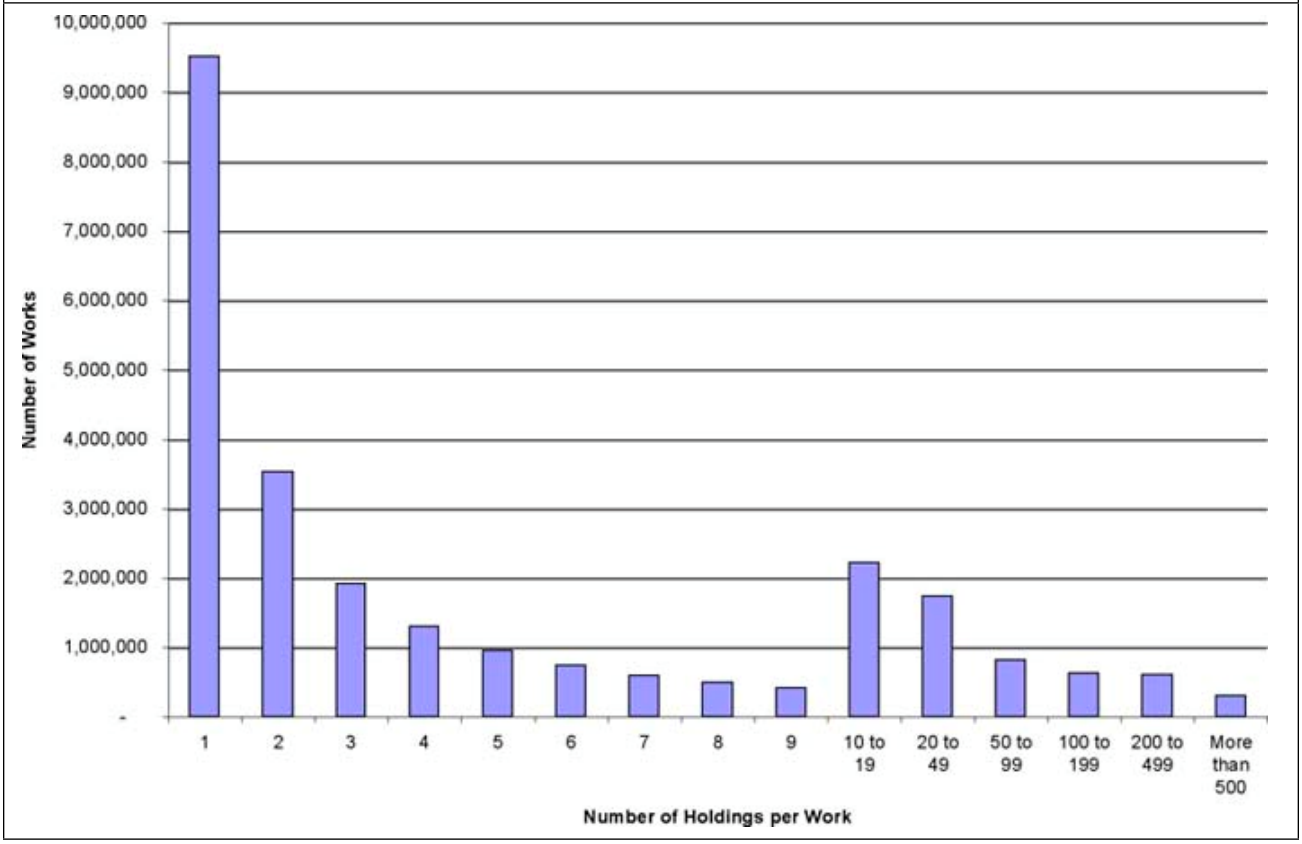


a dataset that ended in January 2005 to show that about half of all library collections were published after 1977, essentially the most recent thirty years at the time of publication, as illustrated in figure 1.[5] A simple analysis of ReCAP's (Research Collection and Preservation Consortium, offsite storage for Columbia, Harvard, and Princeton universities and for the New York Public Library) collection of nearly seventeen million volumes shows similar results.[6] The median publication year of ReCAP's collection among titles published between 1800 and 2004 is 1975. Granted, materials in the ReCAP collection are expected to have lower circulation and may skew older than those materials the libraries retain on-site, possibly explaining the small difference in median age. Holdings from the fifteen years following 2006 make up almost 18% of all of the holdings, moving the median publication date at ReCAP's facility to around 1985.

Research by OCLC has shown that scarcity is common. The collective collections attempt to promise greater depth and continued access to a long tail of low use works that may otherwise be inaccessible. Approximately three quarters of the print book collections held by the Big Ten Academic Alliance (BTAA) members are held by three or fewer BTAA libraries.[7]

Lavoie and Schonfeld also found that there was a significant percentage of unique and scarcely held materials recorded in WorldCat. Many of these materials are locally produced ephemera. About 36% of all titles are uniquely held. The published graph (see figure 2) appears to show that 25% had ten or more copies, including 1.2% that had more than 500 copies.[8]

FIGURE 2
WorldCat Data Showing Number of Holdings per Work
Reproduced with permission from Roger C. Schonfeld and Brian F. Lavoie, "Books without Boundaries: A Brief Tour of the System-Wide Print Book Collection," *The Journal of Electronic Publishing* 9, no. 2 (2006), https://doi.org/10.3998/3336451.0009.208.



There are indeed very widely held materials, but they are a small subset of the corpus held in libraries nationally.

Connaway et al. suggest that the profile of uniquely held materials skews older than overall library collections.[9] They found that the median age book in Vanderbilt University Libraries' collection was published in 1970, while the median publication date of the uniquely held materials was 1928. If this holds true nationally, or at least in academic libraries, it suggests that scarcely held materials have a greater chance of being older and at higher risk than more commonly held materials. It should also be noted Connaway et al. found a disproportionate number of pamphlets among the unique copies. A full 30% of the unique copies were pamphlets, compared to 10% in the full collection. These findings are not dissimilar to those of Lavoie and Schonfeld, that a large percentage of uniquely held materials are local ephemera.

Shared Print and Deaccessioning

Shared print is often seen as a way in which libraries can alleviate pressure on library physical spaces while preserving the scholarly record.[10] Libraries can deaccession local copies based on the presence of a committed copy held elsewhere, reducing costs without eliminating access. Horava describes a future in which libraries relieve themselves of reputations based on the size of their owned collection and focus on what is accessible to their users.[11] Libraries no longer would be gatekeepers to information as much as those who make sense of the ocean of resources. This vision gains the advantage of almost limitless resources at the expense of intangible, impermanent, and unpredictable resources.

The recent growth of shared print programs is extraordinary, with over forty million commitments made by February 2018, essentially five years into the first wave of shared print initiatives.[12] One could imagine—with the growth of shared print, the attraction of deaccessioning responsibly based on external retention commitment, and the commonality of scarcity—how easy it would be to inadvertently reduce the number of copies of a title below the threshold of tolerable risk. Without commonly agreed on retention numbers, it is likely that mistakes will be made, putting titles at a risk of total loss. This risk is especially problematic because of the difficulty in finding replacement copies several years after publication.[13] Agreements and a common idea of risk mitigation can help alleviate some of the uncertainty about loss, but it will require building trust and a broader view of a user base than just those who come through a library's door. "If fewer print materials are available in close proximity to users, it becomes important to ensure convenient discovery and delivery of those materials within new arrangements."[14]

Although it can still be difficult, the sharing of retrospective collections is making cooperative purchasing much more palatable.[15] There is hope that as libraries build trust with each other there will be more opportunities for developing collections collaboratively. Cooperative programs are indeed growing, with many noting cost savings as the primary driver. There is also an additional benefit of collectively extending the collection breadth and expanding what is available to users.[16] Cooperative collecting programs are still gaining steam, and as they do, it is imperative that they consider how many copies will be necessary.

The Monograph Risk Model

The model developed for our study is a generalization of that in Yano et al.,[17] which accounts primarily for physical losses of materials over time, with loss rates being differentiated by

storage conditions. Having identified several other factors contributing to losses, as described below, we developed a mathematical representation of the probability that at least one usable copy remains for each year up to a specified time horizon in the future, given user-specified numbers of books in each user-defined initial condition and storage condition. Once a shared-print program decides on its acceptable probability of at least one usable copy surviving (P1) at a specified point in time in the future (T), it can use the model to aid in searching for the numbers of copies in various initial conditions and various storage conditions on which retention commitments would be needed to achieve P1.

The spreadsheet implementation of this model is flexible and can accommodate a variety of numerical inputs, not the least of which is a usability trajectory that reflects the impact of degradation over time. When this model was first developed, we had not settled on a particular usability curve, so we considered it important to allow the user to input usability estimates that may differ based on the initial condition of the books and on storage conditions.

We assume the value of P1 is selected based on a library's or group of libraries' risk tolerance for losing access to the material during an agreed-upon time horizon. The scale may consist of a handful of libraries or all research libraries in a geographic area or country. A larger group of libraries may naturally desire a higher value of P1 because of the larger aggregate value of availability of the material to the group (versus one library). Consortia may decide to set higher values of P1 for particularly valuable material and/or material that they hope to retain for much longer than the initial planning horizon. Regardless, it is imperative that libraries within a group agree on P1 values and planning horizons; otherwise libraries with lower risk tolerance could unknowingly miss their thresholds if libraries with higher tolerances withdraw books.

The focus of this model is titles, not books. Libraries often discuss preservation and loss in the context of individual items. For the purposes of this project, our team ignored the individual items and concentrated on the combined copies that comprise a title. The model and calculations are for groups of copies that should be considered duplicate intellectual units, which we refer to as a "title." When we talk about loss, we are referring to the total loss of all copies of said title among those held by a shared print consortium. Our work is concerned with minima for preservation; adequate coverage for access is out of this project's scope.

Identifying and Quantifying the Risk Factors

Determining an adequate number of copies that should be retained depends upon a quantitative assessment of the risk factors that contribute to calculating P1 for a given set of retained copies. Then an acceptable risk tolerance must be decided for the target time horizon in view of the cost of retaining the associated number of copies, with the copies possibly held in different storage conditions.

Previous research has identified many factors that can influence risk to usability of books over time, both positively and negatively. The most accessible breakdown of risk factors for heritage collections appears on the Canadian Conservation Institute website, where ten agents of deterioration are listed: physical forces, fire, pests, light, incorrect relative humidity, thieves and vandals, water, pollutants, incorrect temperature, and dissociation.[18] Other guides to risk assessment are also available.[19] We identified four factors as critical for the long-term survivability of monographs: (1) on-shelf probability; (2) bibliographic inaccuracy; (3) accidental physical loss or irreparable damage; and (4) gradual physical deterioration, which depends

on the initial condition and future storage environment.* A discussion of other factors that might impact title availability is provided later in this paper.

Factor 1: On-Shelf Probability

On-shelf probability is the chance that a given item's whereabouts is known. On-shelf probability is calculated only at the time of the initial analysis and provides a stand-in for actual validation at the shelf. We know that research collections have some books that are recorded in the catalog but cannot be located: they are not on the shelf, not checked out to a borrower, not in process, or not otherwise findable. In an ideal world a shared print commitment would begin with a validation check that each book committed to the program can in fact be located; in practice such validations are too labor-intensive to implement. The on-shelf probability factor measures the likelihood that a book selected from the catalog is not available to contribute to the survival of the title in the future.

Our generalized estimate of the probability that a book in the catalog cannot be found on the shelf derives from the EAST Validation Study.[20] The EAST study, which received data from over fifty libraries and assessed over 316,000 books, found that 97% of the items were on the shelf or could be accounted for (or, alternatively expressed, 3% of the items could not be located). Interestingly, the on-shelf rate stayed relatively consistent regardless of publication date across the range from 1850 through 2010, as shown in table 1. The average (and median) on-shelf percentage from 1821 through the end of the study was 97.4% (97.57%).

The fact that the on-shelf percentage remains relatively constant in the data reported in the EAST study lends weight to the theory that overall annual loss is insignificant. (See the discussion on Factor 3, Annual Loss Rate, below.) Known losses would not appear in the on-shelf percentage because those items would be removed from the catalog. The EAST data suggest that unknown disappearances of copies generally occur early in an object's lifecycle, when use is higher, followed by little additional loss of copies as the title ages.

A study at Indiana University found that 2% of materials in the open stacks were missing but went on to say that the material in storage has always been found when requested.[21] Indiana follows a common practice in high-density storage facilities: when items are first processed into storage, staff touch each item and confirm their records. For the libraries that follow these procedures, books are all but guaranteed to be on the shelf with a good quality record. Data from a large swath of research libraries indicates that an average of 1.5% of the total number of volumes circulate each year, and the rate for items in storage is even lower.† Furthermore, anecdotal evidence suggests that only a small fraction of the volumes that do circulate are not returned. If such an item is not returned, presumably the library is aware of it and may attempt to replace it. Even if the library fails to replace a known loss, the prob-

* In our model, all of the physical risks presented in the Canadian Conservation Institute list are merged together as either (3) annual loss rate, which includes accidental physical loss or irreparable damage, or (4) physical deterioration over time. Our risk (2) bibliographic inaccuracy is a form of dissociation, and (1) on-shelf rate is affected by prior actions of thieves and vandals—although such loss could also be an unintended accident.

† The 2019/20 ARL statistics, excluding the public (Boston, New York Public Library) and government (Library of Congress, National Library of Agriculture, National Archive, National Library of Medicine, and Smithsonian) libraries and the Center for Research Libraries, reports that the number of circulations as a percentage of the collection is 1.5% on average. Minimum (Wayne State) = 0.2%, maximum (Brigham Young University) = 5.1%. ("ARL Statistics 2020," Washington DC: Association of Research Libraries, September 9, 2001, https://www.arlstatistics.org/repository.)

ability of loss for a volume held in storage is exceedingly small in the context of the collection as a whole.

To recap: while we recognize that the actual rates may vary widely for different libraries or different collections, we are confident that a rate of 97% on-shelf provides a reasonable estimate for describing broad, generic research library collections in situations where an on-shelf validation step has not been performed. If the copy is held in storage, we use an on-shelf probability rate of 100%. If a library is more or less confident that any given item is on the shelf for a specific collection, a different percentage can be entered into the spreadsheet tool.

In the model, the on-shelf probability rate applies only at the point of analysis, in order to estimate past unknown losses that reduce the number of copies of a title now available to contribute to its survival. Subsequent losses are calculated in (3) annual loss rate and (4) deterioration, as described below.

TABLE 1
Data from the EAST Validation Study, Set Out to Show On-shelf Probability as a Factor of Date of Publication
(Sara Amato and Susan Stearns, East Validation Data, 2018. Unpublished, provided by the authors.)

Pub. Date	Total Number of Books	Present	% Present
<1800	23	21	91.30%
1800–1810	57	54	94.74%
1811–1820	54	50	92.59%
1821–1830	107	105	98.13%
1831–1840	150	145	96.67%
1841-1850	212	203	95.75%
1851–1860	344	334	97.09%
1861–1870	340	331	97.35%
1871–1880	591	579	97.97%
1881–1890	977	951	97.34%
1891–1900	1,719	1,678	97.61%
1901–1910	2,782	2,697	96.94%
1911–1920	3,064	2,986	97.45%
1921–1930	5,948	5,811	97.70%
1931–1940	6,361	6,168	96.97%
1941–1950	8,063	7,815	96.92%
1951–1960	14,763	14,322	97.01%
1961–1970	38,044	36,930	97.07%
1971–1980	41,378	40,223	97.21%
1981–1990	44,399	43,245	97.40%
1991–2000	50,224	49,037	97.64%
2001–2010	40,485	39,635	97.90%

Factor 2: Bibliographic Inaccuracy

Michaels and Neel note that while it is becoming common that libraries are making large-scale retention decisions based purely on metadata, there is concern about the quality of the metadata and the lack of common agreement on loss and risk.

> There are many discussions in the shared print community of how many copies are enough to ensure that the scholarly record is both preserved and accessible. Ensuring that enough copies are retained is a matter of having confidence in the records, but also accepting that there will be a margin of error in the accuracy of holdings statements. If we know that, for example, in most libraries the margin of error is 10%, then we could factor that into how many copies we keep. The difficulty though is in knowing what that percentage of error is so that it can be accounted for across the collective collection. By guessing at a percentage, we risk saving too many or too few copies. There are a few studies that report on error rates that we can look to for guidance; however, more information is needed before broad generalizations can be made.[22]

We recognize that the bibliographic inaccuracy factor is rather specific to this context. Our team concentrated on cases in which the record refers to a discernibly different item than that to which it is attached. Many programs make decisions based on record analysis, not by examining the books themselves—not only when making retention commitments, but also for mass withdrawal decisions. Cases in which there is a difference between the item and the record can result in fewer copies retained than anticipated. For example, if one decides to retain a book based on its record and the item associated with the record is not the desired book, the book will be retained but does not serve its intended role.

We did not consider instances in which poor record quality inhibits good matching. Although poor record quality—such as incomplete records, typos, and missing information—complicate shared print efforts by making record matching difficult, it does not increase the risk of loss. If anything, poor quality records may make it appear that there are more unique titles than actually exist, giving an inaccurate sense of scarcity. Moreover, a falsely unique copy associated with a poor-quality record is, in fact, another retained copy of a different title, if at some point it can be properly identified with that other title.

In the fall of 2019 we performed a study to evaluate bibliographic inaccuracy in this specific context of retention for a shared print collection. Because resource sharing departments look closely at the items they are pulling to ensure they correctly match the request, we asked libraries to track bibliographic errors while processing resource sharing requests in 2019. We received valid results from thirteen libraries for a total of 29,630 requests (each request was for one item) and found an overall 0.1% error rate. This is actually a conservative (high) estimate considering that most of the errors reported did not contribute to confusion about the object in hand. Author and title normalizations were commonly identified as differences. Publication date variances within a year or two were also common and usually did not have separate records in WorldCat. We did not categorize the results by publication date or language, so it is possible that earlier printed books, or books from particular geographical areas, may show higher rates of error.

Michaels and Neel's study of Indiana's collection mostly supports our findings.[23] Although their analysis of catalog records had a different focus than our study, Michaels and Neel performed an in-depth evaluation of Indiana University's collection. They found that 0.54% contained a cataloging error. The record error rate we found is significantly lower than Indiana's findings, but there are reasons for the differences. Over half of the Indiana errors were caused by incorrect home locations. Incorrect barcodes also made up a significant portion of the errors. Incorrect locations and barcodes are inconvenient, but they would not lead a person to identify a substantially different book than what is described in the record. Michaels and Neel only found a handful of these types of discernable catalog errors. Incomplete records were more commonly found than discernable catalog errors that would result in a complete mismatch between the item identified in the record and the book that is physically owned.[24] In practice, incomplete records may not contribute significantly to the risk in our model due to the aforementioned issues where poor-quality records may give an erroneous sense of scarcity. Because of the differences with which Michaels and Neel defined catalog errors, their study does not appear to contradict our 2019 study finding of a 0.1% bibliographic error rate.

Michaels and Neel also found that the confidence in the record accuracy is much higher for items in off-site storage because of common processing practices. Our data suggest that an inaccuracy rate of 0.1% is a conservative (high) estimate for books in library stacks, so a 0.0%

bibliographic inaccuracy rate could appropriately be applied in our model if the collections being analyzed are managed in off-site storage.

To recap: for our analysis we use 0.1% bibliographic inaccuracy—the risk that an error in the bibliographic record would lead to an assumption that the library owns a specific title that it does not in fact own–for items in the stacks, and 0.0% for items in off-site storage. The model is based on the assumption that bibliographic inaccuracy errors affect the calculation for survival only at the point of selection for retention; the record does not become progressively less accurate over time.

Factor 3: Annual Loss Rate

The annual loss rate may be the least intuitive factor of the group. Our experience is that books are lost every year, especially during circulation. In reality, what is lost annually is a very small percentage when put in the context of the collection as a whole. Often a library will replace books that are known to be lost from circulation or irreparably damaged (from a modest water leak, for example) when possible. While we have not found data that suggest a quantifiable number for annual loss, we have included a 0.01% annual loss in the stacks. We feel this number could be the upper echelon of what may be lost annually on top of the probability that an item is not on the shelf at the start of the analysis (see Factor 1). This loss rate, one book lost for each 10,000 books in the collection annually, is most likely much higher than what the average research library experiences each year.[*] Because this is an annual loss rate, the probability of loss compounds over time.

We assume the annual loss for items in storage is near zero,[†] although our model can be modified to accommodate any suitable loss rate. While most materials in high density storage are not in a dark archive, they are generally used at a lower rate than on-site materials. Because of the regular tracking and types of processing performed for materials in storage, it is exceedingly rare that items are not returned or go missing on the shelf. Controlled retrieval also makes libraries more comfortable using protective enclosures, restricting use within a library, or limiting access to supervised areas. Restricted use does not automatically preclude the materials from being included in shared print programs, as some such programs allow the inclusion of materials that can be loaned to a library for use on-site even if a patron cannot take the items home. Many high-density storage facilities report that requested items are always found. Very few items are not returned to storage facilities after use.

Note that the annual loss rate applies regardless of the cause of the loss. This factor lumps together any permanent loss of the whole volume: not returned from circulation, destroyed by water, fire, or earthquake, accidentally dropped down an elevator shaft, or ripped up by vandals. It includes only the lost copies the library does not replace (whether it cannot or chooses not to) and for which a substitute commitment from the shared print consortium is not identified.

---

[*] To aid comprehension of what "one book lost for each 10,000 books in the collection" looks like, consider these two examples. For a modest library holding one million volumes, that loss rate calculates to 100 volumes lost every year that are not replaced. For a large library holding 10 million volumes, that loss rate represents 1,000 volumes lost every year.

[†] Because of its shared print program, ReCAP has tracked incidents at the facility or during circulation since January 2019. As of March 2022, there are nearly 17 million items in ReCAP's facility. While uses were down because of the COVID-19 pandemic, about 509,000 uses occurred between 2019 and March 2022, and ten items were damaged beyond repair. This calculates to about one loss per year for every 5 million items in the facility (0.00006%) or about 1 loss for every 50,000 uses (0.002%).

To recap: for this model we use an annual loss rate of 0.01% for books in open stacks and 0.0% for books in storage. The rate compounds over time (annually); the potential impact of this loss grows larger if the time horizon selected for the preservation of the title is longer.

Factor 4: Physical Deterioration Over Time

In contrast to the annual loss rate above, physical deterioration over time leads to the gradual loss of usability of books. The book may still be on the shelf, but at some point it may become too deteriorated to circulate, read, or scan. Physical deterioration stems from two broad causes: physical wear and tear, typically from use but also from other physical forces, water, pests, and the like; and from chemical deterioration, which is largely influenced by paper composition and storage environment, specifically light, temperature, and humidity. Unlike the On Shelf probability, which reflects a one-time event when lost, physical deterioration happens gradually over time.

We capture physical deterioration via a probability of usability curve. The first point on that curve is an estimate of the probability of usability of the book at the time of analysis (without inspecting it at the shelf) that reflects the sum of the book's experience in the past. Starting at this value, the probability of usability declines with time, forming a probability of usability curve. The shape of that curve depends on storage conditions, including temperature and relative humidity of the storage environment.

Little published research has quantitatively documented the general usability of library collections, especially over time.* While we know paper degrades along a non-linear curve, it is not a given that paper degradation has a direct correlation to book usability. For our model we needed a way to determine the probability that a book is usable without actually inspecting it on the shelf, a process that is too labor intensive to be practical for large collections.

To complete the EAST Validation study, libraries' staff and student workers assessed over 316,000 books on the shelves at over fifty libraries; among other data points, they recorded the books' current condition.[25] Library staff were asked to mark items as being in excellent, acceptable, or poor condition. Poor condition was selected if the book exhibited one of the following criteria:[26]

- Cover
  - □ Obvious water or other damage
  - □ Unattached or loose covers
  - □ Dirty or sticky residue
  - □ Need to wash hands afterwards
  - □ Major fading of color
  - □ Obvious dye discolorations
  - □ Significant markings
- Pages
  - □ Full of markings
  - □ Some pages not legible

* In the 1980s and '90s many research libraries conducted preservation surveys of their collections. A systematic review of these surveys might provide useful data for our model, but there are problematic limitations: many focus specifically on acidic and brittle paper, which may not accurately correlate with usability; many record condition, but do not distinguish between remediable and irremediable damage; most are sampling surveys with a sample size that is too small to derive meaningful conclusions about subsets of material (e.g., by age); and most are one-shot surveys that do not record continued deterioration over time. Most of these surveys are unpublished internal documents.

- □ Torn pages beyond repair
- □ Have folds and creases that cannot be straightened
- □ Obviously missing pages

- Spine
  - □ Spine is broken or almost broken
  - □ Residue from spine glue falls out in flakes
  - □ Pages have come off spine

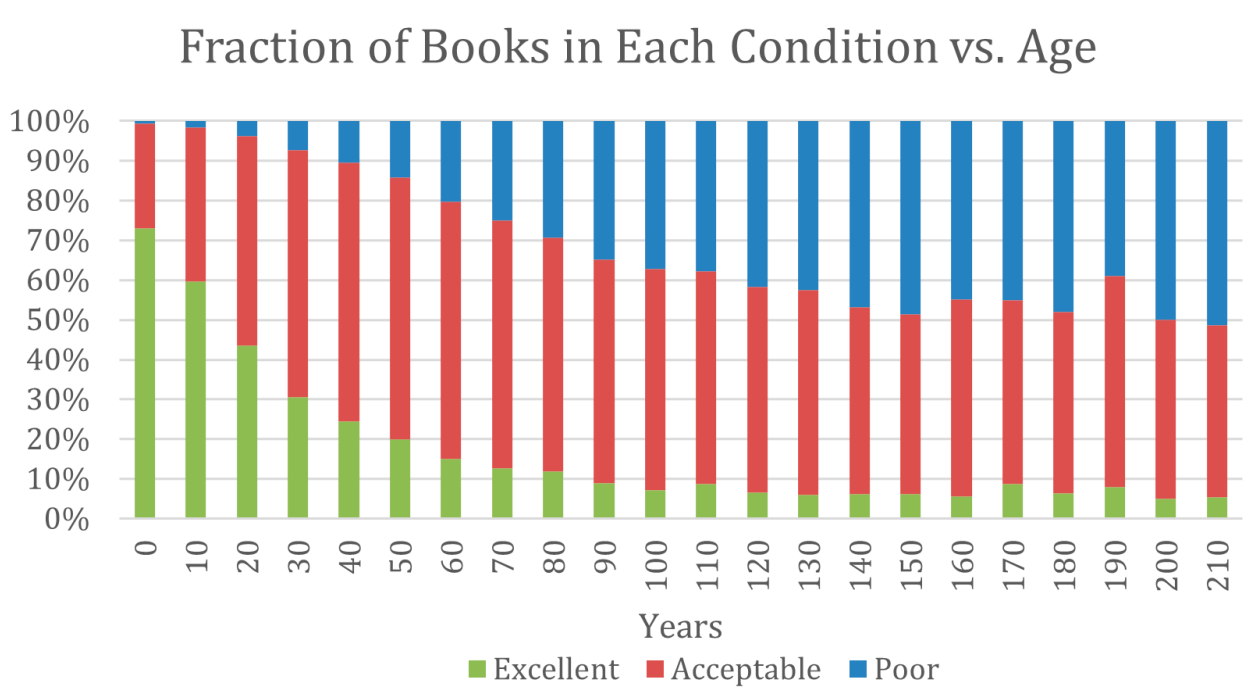
We decided that categorizing the items labeled "poor" in the EAST study as "unusable" offers a reasonable approach for estimating a usability curve over time. The EAST study's definition of poor condition is broader than a practical definition of "unusable." All truly unusable items would be captured in the poor category, but many items in the poor category may well be usable: for example, items requiring the user to wash hands after use, exhibiting dried out glue residue, or other categories describe damage that may be repairable. However, these criteria overall describe volumes that are not fit for circulation as is. Applying these criteria for poor condition as equivalent to "unusable" may be overly cautious, but we decided it was better to be conservative, considering how the definition is applied for our purpose. Usability is subjective and, to the best of our knowledge, there is no other available data that span so many books across multiple libraries.

We used unpublished raw data from the EAST condition assessment to create a graph that shows an approximation of a general (un)usability curve (figure 3). We grouped books in decade-long buckets based on when the books were assessed. Books published from zero to nine years from the assessment were grouped into the zero-year (just published) bucket, books published from ten to nineteen years from the assessment were grouped into the ten-year bucket, and so forth. Consequently, our categorization leads to a slight underestimate of the condition levels of the book, erring on the side of conservatism vis-à-vis risk. The chance that an item is in poor condition increases quickly as it ages until about the 100-year mark, where it starts to level out. About 50% of materials around 150 years old are in poor condition. This curve is similar to paper degradation curves, discussed below, and supports an assumed relationship; although the structural damage may not be visible, loss of molecular weight in paper and paperboard occurs rapidly in the first twenty-five to fifty years, then levels off to a slowly declining long tail.

There are a couple of points worth making. Paper production changed substantially from rag to wood pulp around 1870 and again from acidic to alkaline processes around 1990.[27] These significant historical events could impact the usability curve projecting forward. We think the impact of the change to wood pulp on the usability curve is minimal, considering that the curves are fairly flat for ages of 150 years or greater. The assumption is that the usability curve for books aged 0 to 150 years is captured in this data, after which there is little change.[†] The impact of the change in paper manufacturing from acidic to alkaline processes in the late twentieth century is more challenging to estimate: books of this vintage simply

[†] The number of items in the EAST study drops with increasing age. This is not surprising considering the history of publishing and library collecting, but it does mean that there are relatively few items, less than 500 in total, published in the first half of the nineteenth century included in the study. Because of the decreasing population tested with greater age, the data for each of the early decades has less statistical accuracy. We smoothed the curve after 150 years to 50% usable. The statistics from the EAST data technically rose and fell, although there is no reason to believe that an individual book's usability would ever increase over time. We assume the variability for 160- to 210-year-old books in the EAST data is due to low confidence because of small sample size, and the prevalence of rag paper during that time period.

FIGURE 3
Data from the East Validation Study Set Out to Show the Percentage of Books in Excellent, Acceptable, and Poor Condition, by Age of Book
(Sara Amato and Susan Stearns, "Documenting the Stewardship of Libraries: The Eastern Academic Scholars' Trust Validation Sample Studies," Collaborative Librarianship 10, no. 3.4, 2018, https://digitalcommons.du.edu/collaborativelibrarianship/vol10/iss3/4)



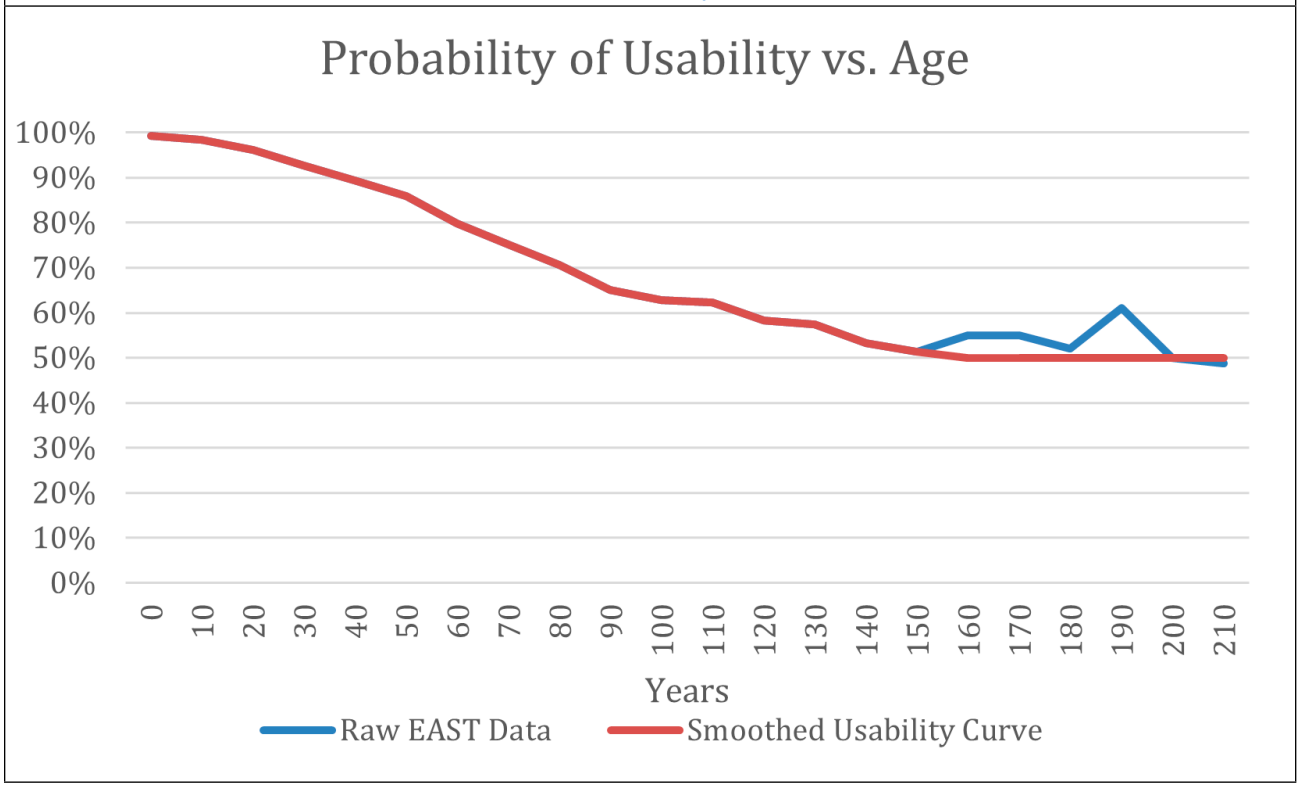
have not been around long enough yet to develop a reliable usability curve. Our curve may overestimate the future degradation for these books, but overall that error will tend to reduce risk for these titles.

Using the data from the EAST validation study, we estimated the probability of usability as the ratio of the number of items in acceptable and excellent condition in an age interval to the total number of items in that age interval (figure 4). From these estimates for each decade-long age interval, we constructed usability curves that reflect the degradation of usability over time, applying some smoothing to eliminate irregularities.

Tétreault, Bégin, Paris-Lacombe, and Dupont provide more nuances on paper decay.[28] The degradation of paper fibers starts essentially at production at a fairly rapid rate. After a period of time, the degradation curve levels out to a long tail. One can clearly see this shape in their graph in figure 5. The paper used in their study was Whatman No. 1 filter, a standard paper frequently used in paper research; it was artificially aged by subjecting it to elevated temperature and humidity. The degradation curve for 128-year-old paper still has some curvature, but the 200-year-paper shows a much more gradual decline over the next four hundred years. These curves are similar to the one formed by the EAST usability data suggesting that paper degradation and usability are indeed correlated.

Tétreault et al. go on to look at the impact of different temperatures and humidity condition for artificial aging (figure 6). Lower temperature and humidity draw out the degradation over a longer period of time; in other words, lower temperature and humidity slows down the process by reducing the chemical reaction rate.

FIGURE 4
Data from the EAST Validation Study Set Out to Show Probability of Usability (Expressed as a Percentage) as a Function of Age
(Sara Amato and Susan Stearns, "Documenting the Stewardship of Libraries: The Eastern Academic Scholars' Trust Validation Sample Studies," Collaborative Librarianship 10, no. 3.4, 2018, https://digitalcommons.du.edu/collaborativelibrarianship/vol10/iss3/4)

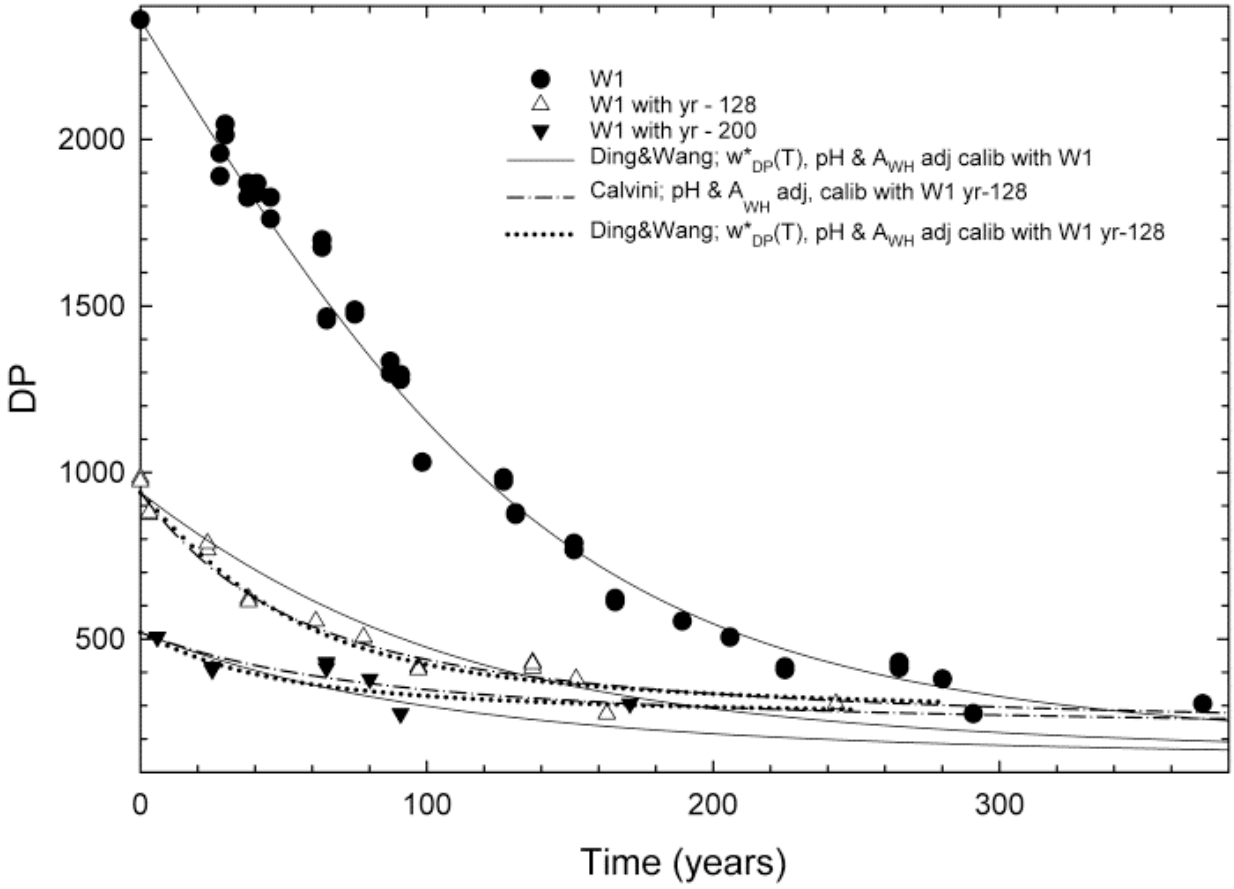


The effects of temperature and humidity on the chemical deterioration of library collections are well researched and documented. Both temperature and humidity significantly influence the decay curve. In general, higher temperatures and humidities will cause paper to degrade more quickly than lower temperatures and humidities. Michalski gave a clear general rule of thumb that each 9˚F (5˚C) drop in temperature doubles the expected life of many library materials, including acidic paper. Independently halving the humidity more than doubles the expected life of these materials.[29] The combined effect of reducing both temperature and humidity more than quadruples the expected life. These estimates are reinforced by the Image Permanence Institute's Dew Point Calculator (http://www.dpcalc.org/), an interactive tool where one can select different combinations of temperature and humidity and see the impact on four preservation metrics: natural aging, mechanical damage, mold risk, and metal corrosion.

In our analysis we need to account for both the current usability of a book at the time of analysis and the decline in usability in the future, which depends upon the future storage environment. We consider two storage environments: library stacks (assumed at ~72˚F, ~45% RH) and lower-temperature, lower-humidity conditions typical of offsite storage (assumed at ~55˚F, ~35% RH). Because storage facilities with high-quality environmental conditions are relatively new (Harvard opened the first purpose-built facility in 1986), we assume the current condition of books is as if they had been held in the stacks since publication. We determine the chance that a book is usable at the time of analysis by finding the book's cur-

FIGURE 5
Trend of Decay Curve Models Based on Experimental Data of Whatman 1 Paper Aged in Closed Tubes

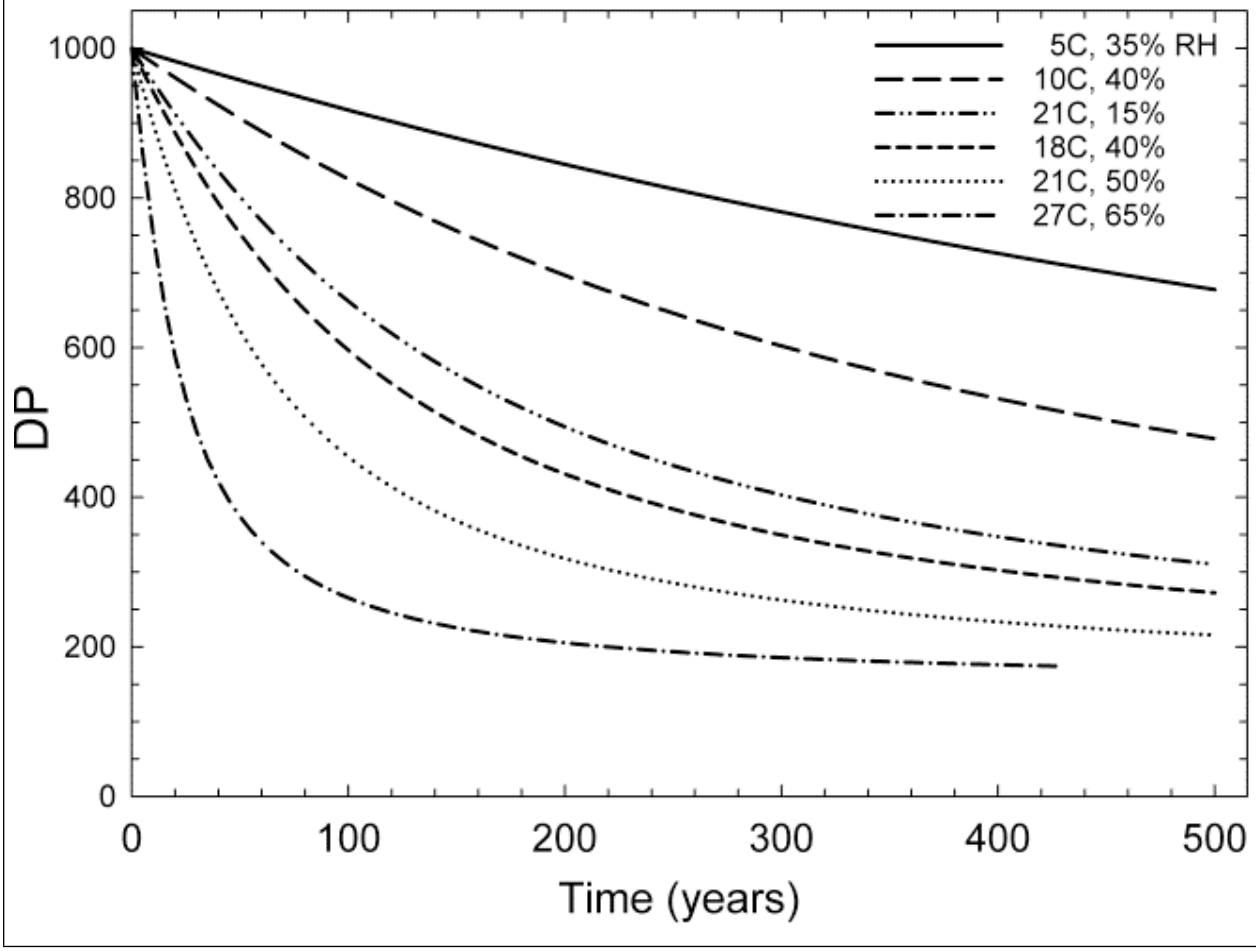
In the graph, W1 refers to Whatman No. 1 paper, DP stands for degree of polymerization and A_{WH} is a factor that reflects the impact of hydrogen ions on paper aging. Reproduced with permission from J. Tétreault et al., "Modelling Considerations for the Degradation of Cellulosic Paper," *Cellulose* 26, no. 3 (February 1, 2019): 2013–33, https://doi.org/10.1007/s10570-018-2156-x.



rent age on the EAST usability curve. Looking into the future, our usability curve splits into two alternative paths. If the book is held in the stacks, its usability continues along the same curve as estimated from the EAST data. If, however, the book is held in storage, the book's usability curve starts (again, at the time of analysis) at the estimated usability for the book of that age based on the EAST data but has a slower rate of decline. We assume that if a storage environment decreases the chemical degradation by a quarter, the degradation over the next 100 years will be equal to the effect of aging in the stacks for 25 years. The environmental improvement between storage and stacks more than quadruples the life expectancy of the book. The conditions may differ in real life situations, although any 17°F and 10% RH difference between stacks and storage should lead to similar results. The usability curves for books held in storage in the future depend upon the current age of the book; examples of a few curves are shown in figure 7.

We regard usability curves derived from the EAST data as worst-case scenarios because storage conditions going forward are expected to be significantly better than they were even forty or fifty years ago–even environmental conditions in stacks are far better. Therefore, the

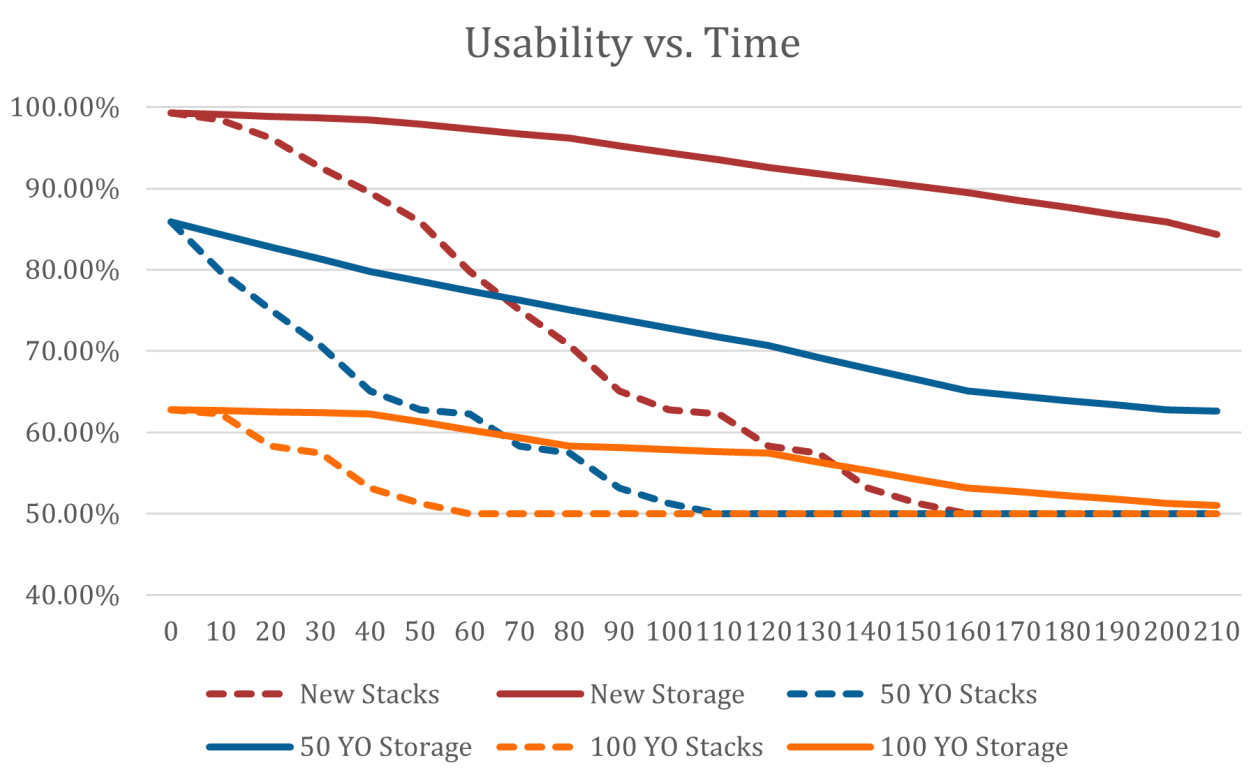
FIGURE 6
Calvini Model Simulation: Decay Curve for Acidic W1 Papers in Stack Under Different Hygrothermal Conditions
Reproduced with permission from J. Tétreault et al., "Modelling Considerations for the Degradation of Cellulosic Paper," *Cellulose* 26, no. 3 (February 1, 2019): 2013–33, https://doi.org/10.1007/s10570-018-2156-x.



usability curves that we utilize are also conservative (understating the probability of usability). Also, we acknowledge that the intended use of the physical items could influence the definition of usability and therefore also the specification of the probability that the item is usable at various points in time in the future. For example, a book containing only text may be usable, provided that the reader can make out the letters, but a book containing maps that is in a similar state of deterioration may be regarded as unusable.

To recap: our measurement for physical deterioration combines a calculation of the probability of usability at the time of analysis, based on the age of the book at that time, with estimates of further deterioration in the future, based on a combination of age and storage environment. The calculation for future decay is differentiated according to two options for storage environment: (1) typical conditions in library stacks and (2) typical conditions in off-site storage. For the model in the next section, we consider two commonly occurring types of storage, open library stacks (72° F and 45% RH) and typical off-site storage (55° F and 35% RH) without circulation restrictions. We note that our spreadsheet model is flexible and allows a user to input different usability curves, if needed.

FIGURE 7
Usability Vs. Time for New, 50-Year-Old, and 100-Year-Old Books in Stacks and Storage



Applying the Model

Calculating P1

We next provide a simple example to explain the essence of the calculations to determine P1 at a designated future point in time, which we refer to as T in the discussion below. This is followed by the development of a general formula. Recall that a book will not exist and be usable at some future point in time if any of the following is true: (1) it is not on the shelf at the time of analysis and therefore assumed not to be on the shelf at time T; (2) the bibliographic record is inaccurate at the time of analysis, so what we believe to be a specific book is, in fact, another book; (3) the book is lost between the time of analysis and time T, which depends upon storage security and the elapsed time; (4) given the condition of the book at the time of analysis and the degradation until time T, the book is anticipated to be unusable at that point, which depends upon the quality of storage conditions and the elapsed time.

The probability that the book exists and is usable at time T is the probability that none of these conditions is true. We assume that each of these categories of conditions occurs independently of one another, which may not be technically accurate. For example, the on-shelf rate and the bibliographic inaccuracy may be correlated, with both rates being higher for older books. However, the impact of loss and degradation tends to be far greater than the on-shelf rate or bibliographic inaccuracy for time durations of interest, e.g., 50 or 100 years, as we explain in more detail later. If one is concerned about potential adverse correlations, one option is to overstate the risks when choosing numbers, which will lead to conservative choices about the number of book copies to retain.

We now present a numerical example. To keep the exposition simple, we will assume T=50 years is selected in advance and all numerical values are selected consistently with that T. The scenario for our example, a group of five copies of one title, is described below, with numerical values chosen so as to avoid confusion:

- On-shelf probability: 97% in library stacks and 100% in storage
- Bibliographic inaccuracy rate: 0.1% in library stacks and 0.0% in storage
- Three books in very good condition (90% probability of being usable) and stored in off-site storage with good environmental controls; physical loss probability of essentially 0.0% per year and degradation (reduction in probability of usability) down to 70% probability of being usable at time T.
- Two books in excellent condition (100% probability of being usable) and stored in library stacks (less secure storage and weaker environmental controls); physical loss probability of 0.01% per year and degradation down to 60% probability of being usable at time T.

Although it would not be common to have copies of the same title that we know to be in different conditions (calculated on the basis of age) at the start of the planning horizon, if there were two printings of the title, with the original occurring about twenty-five years ago and another occurring very recently (and assuming we do not treat them as distinct titles), then we would expect copies of the original printing to be in very good condition and the recently printed copies to be in excellent condition. However, for the purposes of our example, what is important to distinguish is that the two sets of books will have different levels of usability at time T. This may be a consequence of starting out in different conditions, being stored in different conditions in the future, or both.

Considering only the physical loss for a book stored in library stacks, the probability that it is not lost after one year is 1 – 0.0001, so the probability that it still exists at T = 50 is:

$$(1\text{-}0.0001)^{50} = 0.995$$

Incorporating the other factors for one of the books stored in library stacks, the probability that it exists and is usable at T = 50 is:

$$(0.97) * (1\text{-}0.001) * [(1\text{-}0.0001)^{50}] * (0.6) = 0.5785 \text{ or } 57.85\%.$$

The expression in the first set of parentheses represents the probability that the book is initially on the shelf and that in the second set of parentheses is the probability that the bibliographic record is accurate. The expression in square brackets is the probability that the book has not been lost after 50 years, and 0.6 is the probability the book is usable at T = 50.

Analogous calculations for a book stored in off-site storage with good environmental controls is:

$$(1.00) * (1\text{-}0.000) * [\ (1\text{-}0)^{50}] * (0.7) = 0.70 \text{ or } 70.0\%.$$

Each book has a probability of 70.0% or 57.85% (depending on storage type) of existing and being usable at T = 50. The probability that at least one of them exists and is usable at T = 50 is simply 1 minus the probability that all five of them do not exist and/or are not usable at T = 50, which is equal to:

$1 - [(1 - 0.70)^3] * [(1 - 0.5785)^2] = 0.9947$ or 99.47%.

In essence, we determine a value such as 0.70 or 0.5785 for each individual book at time *T*, and from these values calculate the probability that at least one of them survives and is usable at that time.

We have chosen realistic or somewhat realistic numerical values for this example. Even from this simple example, it is clear that the decline in the probability of usability is a dominant factor.

We now develop a general formula for P1. To do so, we distinguish copies of a book title by the combination of their initial probability of usability (at the time of analysis) and their associated storage option, which we refer to as a type, indexed by *i*. Given this information on type *i*, we can read the following value from a table or graph of the corresponding usability curve:

u_{iT} = probability that any given copy of a book of type *i* is usable at time *T*, assuming that it exists.

Additional notation is defined as follows:

α = on-shelf probability

β = bibliographic inaccuracy rate

γ = annual loss rate

n_i = number of copies of type *i*

N = number of types

T = retention horizon.

We can now express $p_{iT}p_{iT}$= the probability that any given copy of type *i* exists and is usable at time *T* as:

$$p_{iT} = \alpha \cdot (1 - \beta) \cdot [(1 - \gamma)^T]\, u_{iT}$$

Then, P1 can be expressed as follows:

$$\text{P1} = 1 - [(1 - p_{1T})^{n_1}] \cdot [(1 - p_{2T})^{n_2}] \cdots$$

$$1 - [(1 - p_{1T})^{n_1}] \cdot [(1 - p_{2T})^{n_2}] \cdots [(1 - p_{NT})^{n_N}][(1 - p_{NT})^{n_N}]$$

Note that each bracketed term is the probability that no copies of type *i* exist and are usable at time *T*, and the product of the bracketed terms is the probability that no copies of any type exist and are usable at time *T*.

Our spreadsheet version of the model accommodates books with different initial usability estimates (probabilities) and different degradation trajectories. In the spreadsheet, we calculate a trajectory of these values for user-selected time grid points (e.g., multiples of a decade) up to time *T*.

Running the Model and Interpreting the Results

There are two major dimensions of specifying acceptable loss. The first is the selected time horizon, *T*; the second is the acceptable probability of loss over that time horizon, P1. We discuss each of these in turn. We selected 50, 100, 150, and 200-year time horizons for analysis because collection managers may desire different planning horizons for books of different ages or importance. The spreadsheet model is flexible and allows consideration of any user-selected time horizon, but we decided not to model horizons shorter than 50 years because we wanted to avoid being too myopic and thereby understating true retention requirements.

Choosing the minimum acceptable probability of survival, P1, is also critical. The spreadsheet model will calculate the probability of at least one surviving copy at the end of the specified time horizon using the given inputs (including the number of copies in each type of storage), but determination of whether that probability is acceptable rests with the collection decision-makers. We report results for a probability of survival of at least one copy (P1) of 99.8%, i.e., a loss of 1 title in 500. We chose this value because it specifies a high level of risk protection while avoiding, in most cases, the need to retain additional copies that provide very small marginal returns in terms of risk reduction.

Setting acceptable values of P1 and time horizon is necessary but the value of P1 says little about what occurs after we reach the horizon. Once the selected horizon is reached, there will be fewer copies and books will have aged. Libraries may not be able to achieve the same value of P1 for additional decades if they pare down to the minimum number of copies needed to reach their first milestone. After all, not only will some titles be fully lost at this point, but also the rest will have at least, but possibly no more than, one usable copy.

The calculations in the results that follow are based on the following scenario, where the selected values are based on studies described earlier:

- on-shelf probability: 97% for books in the stacks and 100% for books in storage
- bibliographic inaccuracy: 0.1% for books in the stacks and 0% for books in storage
- annual loss rate: 0.01% for books in the stacks and 0% for books in storage
- usability trajectories as described in the subsection on Physical Deterioration over Time.

All of the parameters mentioned above were selected to be conservative, i.e., loss or unavailability rates are slightly overstated relative to what available studies indicate. However, they do not explicitly account for rare events that may lead to catastrophic losses. In the subsection on Natural Disasters and Geographic Diversification later in the paper, we explain how such events can be accounted for in an approximate way, and other measures that can be taken to mitigate the effects of such rare events.

We used the model to identify the smallest number of copies of a title needed to ensure a 99.8% probability of survival of at least one usable copy for books of different initial ages at the beginning of the scenario (new, 50 years old, and 100 years old), assuming they were stored in the stacks up to year zero and that future storage of all copies would occur in either off-site storage or library stacks (not a mixture). As mentioned above, we utilized the probability of usability curves as described in the subsection on Physical Deterioration over Time. We note that data on the condition of books older than 150 years is sparse; we have specified degradation trajectories that we believe are conservative, reflecting a smaller probability of usability than what we anticipate will be true at each time grid point.

With the aforementioned estimates, using a range of retention quantities, we calculated survival probability of the title for books of age 0, 50, and 100 years at the time of analysis and

for retention horizons of 50, 100, 150, and 200 years. These values are shown in Appendix A. From these values, we gleaned the minimum retention requirement, i.e., the smallest number of copies that would ensure a 99.8% probability of survival for each scenario (age of book, storage environment, and retention horizon). These values are shown graphically in figure 8 and numerically in table 2.

We make a few observations based on the results shown in the graph. First, if the books are held in the stacks and the retention horizon is long (150 years or more), the current age of the book has little impact on the required number of copies; 10 copies will be needed. Second, if the copies are held in storage, then the minimum retention quantity is sensitive to both the age of the book and the retention horizon.

Third, the reduction in the number of copies needed if the books are moved from the stacks to storage is often greater for newer books, particularly for retention horizons of 100 years or more, because the payoff for higher-quality storage conditions is significantly greater during an item's first few decades of life. This may, at first, seem counterintuitive

FIGURE 8
Number of Books Required to Achieve 99.8% Probability of Survival of at Least One Usable Copy

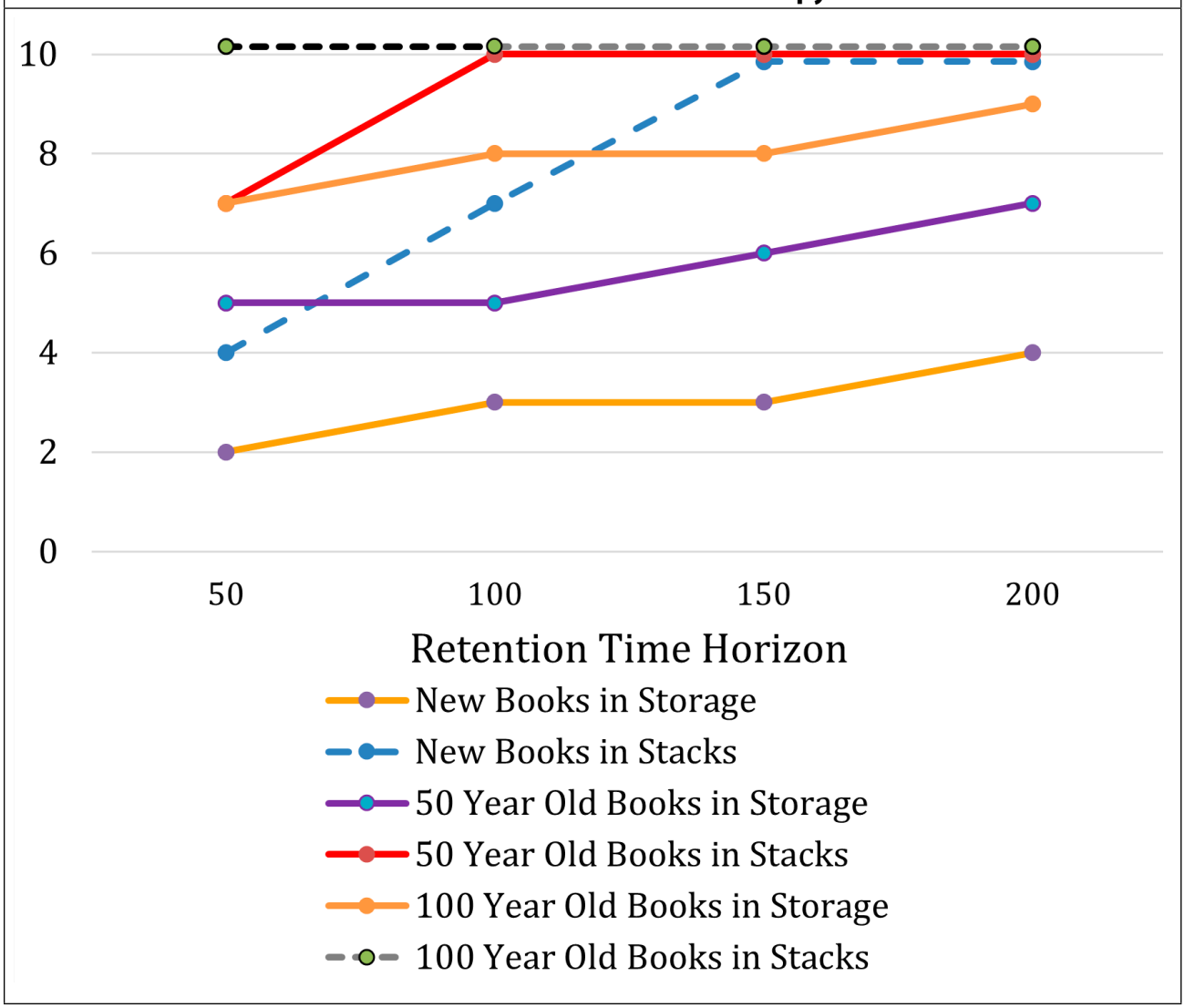


TABLE 2
Number of Books Required to Achieve 99.8% Probability of Survival of at Least One Usable Copy

Storage Condition		Stacks			Storage		
Age of Books at Start, in Years		0	50	100	0	50	100
Retention Horizon at Start, in Years	50	4	7	10	2	5	7
	100	7	10	10	3	5	8
	150	10	10	10	3	6	8
	200	10	10	10	4	7	9

because one might be inclined to store older books in higher-quality storage. This strategy is sensible for scarce titles because the number of available copies is already so small—perhaps only one—and special preservation efforts may be needed to protect the title from complete loss. However, books that are already 100 years old or more have already experienced a large portion of the possible chemical deterioration. The impact of placing those books in higher-quality storage is diminished compared to placing new books in high-quality storage, which significantly slows down the deterioration at a time of a book's life when most of the deterioration occurs. Although placing old books in high-quality versus lower-quality storage does not help as much as we might hope, the analogous reduction in the retention requirements for old books is still significant, up to five copies (for fifty-year-old books and a 100-year retention horizon). These reductions may be quite meaningful, especially when the number of extant copies is small.

Up to this point, we have considered books of several different ages and several retention horizons, and from our results decision-makers can identify minimum retention numbers that apply to their situation. We now take a different perspective and ask how many copies are needed if the goal is to reach P1 of 99.8% at the time a title reaches its 150th birthday, which we think is a reasonable way to view the decision problem, as such a goal is not too myopic and is within the realm of feasibility. From the information in figure 8, we find the following minimum requirements:

- For books in the stacks with a retention horizon of 150 years less the age of the book (i.e., if the collection manager wishes to achieve a high survival probability of each title to age 150): ten copies would be required whether one is concerned with 100-year-old books and a 50-year horizon or 50-year-old books and a 100-year horizon or new books and a 150-year horizon.
- For books in storage with a retention horizon of 150 years less the age of the book: three for new books, five for 50-year-old books and seven for 100-year-old books.
- A mix of books in stacks and in storage could suffice. For example, if there are fewer than ten copies of a book available but not all can be moved to storage, a mixed-storage arrangement may provide adequate coverage. For example, two copies in storage and six copies in the stacks will suffice for 50-year-old books.

Because retention quantities must be integers, some changes, particularly small changes, in parameters may have no impact on the minimum retention quantity. In our model, there are some factors that inflict a one-time "hit" on the survival probability of a copy of a book, namely the on-shelf probability and the probability of bibliographic inaccuracy. The annual loss rate, on the other hand, has a compounding effect over time. With the small, yet pessi-

mistic, rate of 0.01%, the effect of compounding is modest, but one should be careful about the effects of compounding over long horizons even if the annual loss rate is 0.1%. The most important factor in our model, however, is the probability of usability, which, as an example, would be only 63% at a time horizon fifty years from now for a book that is currently fifty years old and will be held in the stacks. This effect dominates that of the one-time effects of bibliographic inaccuracy and not-on-shelf probability.

Factors Not Included in the Model

Above we discussed the risk factors that are included in our calculation of the probability of survival of book titles. Other risks also tend to be raised both in the literature and informally when exploring the topic of factors impacting book retention. In order to gain a better understanding of these additional factors of concern for librarians, the authors both reviewed the literature and met with several advisory panels in fall 2021 and winter 2022. It is important to tease out these "top of mind" issues in order to examine their likelihood of impacting our targets and our ability to calculate the probability of survival of a title in a way that incorporates all important factors.

Natural Disasters and Geographic Distribution

When assessing potential risks to books, we decided to include in our model only factors that could be quantified from available data. Natural disasters are a good example of a risk that is challenging to quantify in this context. Major disasters do occur in libraries, but not frequently enough to reduce the overall number of copies of a given title in a significant way. (Natural disasters may be devastating to collections of unique items, but that scenario is outside the scope of this model.) Moreover, while several organizations track natural disasters that impact cultural organizations—such as the Federal Emergency Management Agency (FEMA), the Heritage Emergency National Task Force (HENTF), and the American Institute for Conservation National Heritage Responders (NHR)—they record incidents in libraries, archives, and museums but not the numbers of volumes lost. Anecdotal reports from insurers of library collections suggest that losses exceeding deductibles are very rare; in other words, that losses have been low when put in context of a library's entire collection.

Still, there may be a good reason to account for the likely increase in adverse weather events. Climate change is predicted to increase the number and severity of weather events.[30] It is difficult, however, to estimate the chance that these events will lead to a significantly greater number of lost copies stored in specific locations. The eastern and southern coastal regions of the U.S. have seen an increase in hurricanes, and are predicted to experience more. The West Coast is prone to earthquakes and is overdue for a large one. About 55–60% of U.S. library collections are held in these regions.[31]

While it is unlikely that a single event, or even multiple events in a given time period, would destroy all of the collections in an area, it is possible that significant losses will occur over time. The impact of events could be higher as collections are consolidated in relatively few locations. Even if we assume that libraries will continue to adapt and improve preventive measures and preservation emergency response capabilities, there may be a threshold that calls for major revision of how shared print archives are planned and managed—including a thorough revision of the numerical values that serve as input to the calculations of our cur-

rent model and possibly an extension of the model to account for catastrophic risks explicitly. Meanwhile, striving for geographic distribution for storage of the retained copies of a title is a recommended best practice. Options might include intentionally making retention commitments for copies held at different institutions, at different campuses of the same institution, or even at different locations within one institution.

To account for natural disasters in the current model, the numerical loss rate that we (and other users of the tool) specify may include additional buffers, i.e., increasing loss rates to account for these additional risk factors. When doing so, one may be tempted to focus on infrequent catastrophic events, but even catastrophic events such as major earthquakes have a low probability of leading to any losses, sometimes due to loss mitigation efforts, so it is important to use realistic numbers. Because our spreadsheet model is intended to be used for time horizons of at least 25 years and more typically 50 or 100 years, one way to incorporate the effects of disasters that occur sporadically is to estimate the probability that a typical item will be lost over a horizon of 25, 50, or 100 years owing to all types of sporadic disasters that one wishes to include, and convert these to annual rates. For example, suppose it has been estimated that there is a 75% chance of a 7.0 magnitude earthquake during the next thirty years on a certain fault in California. This translates to 2.5% per year (or perhaps 3% or 4% if one wants to add a cushion). We would multiply this by the probability that such an earthquake would cause a typical book to be lost or irreparably damaged, which might be (for the sake of this example) 0.1% for a library situated on top of the fault (a loss of 1 book out of 1,000 in that library). We could then attribute a loss rate of 2.5% x 0.1% = 0.0025% to such earthquakes and repeat this process for other types of sporadic disasters. We can then add the annual loss rates from sporadic events to those due to more common, regularly occurring events, and finally use an adjusted annual loss rate that considers both common and sporadic events. If the annualized loss rate due to sporadic events is difficult to estimate, then one can include a buffer to be conservative. If the frequency or severity of sporadic events differs widely from one storage location to another, it is, in principle, possible to separate copies of books into finer-granularity groups and apply appropriate annual loss rates to each.

Impact of Retention Agreement/Accidental Withdrawal

The impacts of retention agreements and accidental withdrawal are also hard to measure. We recognize the significant importance of retention agreements and their implementation—for example, whether these commitments are made public, whether there are clear shared print memoranda of understanding, and whether withdrawal intentions of the parties to the agreements are specified.[32] But how can one measure the effect of retention agreements (or the lack of them) and put a number to it? Likewise, withdrawal decisions are intentional, but one cannot reasonably assign a value to the probability that a copy would be accidentally withdrawn. For these reasons, our model is based on the assumption that copies will be kept unless there is unintentional loss or irreparable damage. For all intents and purposes, the recommended number of copies from this model is synonymous with a minimum number of required retention commitments.

Duplicate and Unique Copies

Much has been written about how books should be compared when considering them as potential duplicates. In 2015 and 2016 Jacob Nadal, Andrew Stauffer, and Mike Garabedian

participated in a friendly debate in the pages of *Against the Grain*. Nadal started the thread describing the pressures for withdrawals and a methodology that he investigated at UCLA.[33] Stauffer eloquently continued the conversation, raising concerns about differences in copies that are not described in Machine-Readable Cataloging (MARC) records. As he says: "Any 'fool' can look at a spreadsheet of 500+ identical pieces of metadata and call the books they reference 'duplicates.'"[34] Garabedian concludes the debate with an experiment where he estimates that it takes about ninety seconds to fetch and record information about books with the intent of comparing their conditions. He found that 31% had "paratextual" value such as original dust jackets, original paperback bindings, or facsimile paperback bindings.[35]

The debate over what Teper calls "sameness" is an important one, and one that she extends to differences in bibliographic data within the MARC record.[36] These are issues that are hotly debated and worth time and attention, but our intent in the present study is something different. Our model was not targeted to situations in which an individual item has significant artifactual value, such as significant marginalia or a distinct binding, that effectively makes it unique. We are attempting to provide guidance on what to do once a group of items has been identified as "the same." Copies that need to be considered different are not part of such a group, but rather fully independent "things."

Our model could be applied to unique items by treating such items as if each is a different title, but the utility of doing so is questionable. It is unlikely that two or more effectively identical copies exist, so the only decision may be whether to retain the unique copy and in what types of storage conditions. There could also be cases where an item could be both part of a group and unique. For example, a signed book has all of the text of the original, but also carries added value in the form of the signature. That copy may serve a dual purpose, an example of the generic title while simultaneously being a unique artifact. We note that although these unusual copies would not be regarded as part of the pool considered to aid in retention of the "standard" version of a book, they may nevertheless provide another backup of the contents.

Digital Copies

The focus of this study is on physical copies; the existence or not of a digital surrogate has no influence on the calculation whether or not at least one physical copy will remain at the selected time horizon. On a broader scale, however, the existence of a digital copy may influence the management decision as to how much risk to the print title is acceptable.

There are several important reasons for retaining access to a print copy even after a digital surrogate has been created: as a source of information not captured (or not captured adequately) in the digital copy; as a source for rescanning if the digital copy is lost; as historical evidence of the original publication; to accommodate researcher preferences for reading and use; and for artifactual evidence that could be difficult to capture digitally.

Moreover, page-by-page validation of digitized copies is complex. Although significant improvements in validation have been implemented since the first forays into mass digitization projects, errors do slip through.[37] Retrospective validation of digital copies at scale by libraries is usually too resource-intensive to contemplate, so errors are discovered randomly when the title is accessed. One purpose of our model is to help libraries avoid a situation where no print masters of a title remain—regardless of whether a digital surrogate exists.

The existence of a digital surrogate may reduce physical wear and tear on the print original. While there is early research that says digitization may increase use of materials, especially

for special collections,[38] more recent analysis finds that the presence of digitized versions reduces circulation.[39] Interestingly, digitization may simultaneously reduce circulation and increase physical sales. Nagaraj & Reimers found that increased sales were most prominent with low use or little-known materials because of enhanced discovery. Well known or highly used materials did not experience increased sales.[40]

Discussion

Developing the model and running analyses of examples with books of different ages and characteristics cast light on several questions about shared print collections and retention that we discuss more fully here.

Not Enough Copies

Inherent in the concept of the on-shelf probability factor described above is the acknowledgment that some titles are already lost. Inherent in the annual loss rate and deterioration factors is the alert that the longer libraries wait to make decisions about retentions, the more titles will be lost.

Even if copies exist at the time of initial analysis, it is expected that some titles may not have enough copies available in libraries to meet the target probability of at least one usable copy remaining at the selected time horizon. In fact, it is precisely those books that need the most retained copies, older titles, that will have the fewest copies available.

In situations where it is not possible to combat the risk of loss by adding extra copies to the pool, other risk mitigation strategies must be deployed. Many of these are already established practice in research libraries: remove older and rare materials to offsite storage or a special collection, validate the existence of copies and the cataloging, apply enhanced preservation measures to stabilize the items and prevent future damage, and exercise tighter controls on circulation, or ensure the title has been adequately and completely captured digitally. Even in situations where it is possible to add extra copies to the pool, libraries and consortia that are making related decisions need to consider costs holistically: the cost of retaining copies of a book (e.g., storage space) and the equipment and energy needed to provide different quality levels of storage conditions, and the technology and labor to maintain better circulation controls. Such investments may lead to needing significantly fewer copies. The best strategy may be different for different portions of the collection.

Ideally, every title held in the national collective collection would be secured with an adequate number of retention commitments. It was estimated that in 2005 there were thirty-two million print book titles in WorldCat.[41] Many titles will not have enough coverage to attain the desired probability of survival of at least one copy until a designated time horizon, even if every copy is committed. Solving the problem of what to do with titles that cannot reach that target will require strategies to mitigate information loss. While outside the scope of this paper, further work should be considered such as digitization or other practices.

Dark Storage

One recurring proposal for the long-term preservation of books is the creation of a dark archive: record-validated copies are placed in non-circulating storage for the entire desired time horizon and are removed only for special circumstances such as to correct digital surrogates.[42] This strategy reduces or eliminates many of the risks. The validation process eliminates the risk that the item might not be on the shelf and substantially decreases the probability that the

bibliographic record is inaccurate. Eliminating circulation reduces the risk of loss. An inspection process would allow for an exact determination of initial condition and usability as opposed to an estimated probability that an item is usable. Future chemical deterioration in a dark archive is similar to light archives as it is mostly influenced by the environmental conditions.

When choosing among alternative storage conditions to include in our analyses, we considered dark storage as an option, but from the discussion of risk factors above, it is evident that there are few differences between dark storage and a circulating collection in an environmentally optimized storage facility. Bibliographic inaccuracy would be 0% in both cases, as would the not-on-shelf rate. Likewise, assuming that the dark storage is in typical off-site storage conditions, the usability curves would be the same as for the storage conditions we used for the model. Given that we have assumed that the annual loss rate in off-site circulating storage is statistically zero (we use zero as an approximation), the annual loss rate for dark storage would be the same.[*] In view of the fact that all of the risks are the same or essentially the same for dark and off-site circulating storage, we include off-site circulating storage explicitly, but the same results would apply to dark storage.

We note that circulation can in fact help the security of the collection. In closed stack and storage facilities, small pockets of damage—a water leak, pest incursions—are typically found when staff go into the stacks to retrieve items for circulation, and the damage can be mitigated before the problem spreads.

If one is interested in considering dark storage *at especially low temperature and humidity*, appropriate usability curves can be entered into the spreadsheet tool and minimum retention numbers recalculated.

A true dark archive may offer greater gains for serials than for monographs, especially since there have been concerted efforts to create digital backfiles of serial runs. Past use patterns for print serials make missing pages and physical damage to heavy bound volumes more common; serials benefit more from a thorough validation process (article or page level validation) at the point of transfer into the archive. Going forward, use for print serials—if they are used—is more likely to involve scanning (by library staff) of individual articles than circulation of whole volumes.

Specialized Subcollections

For the purposes of this paper, we utilized broad-based, generic averages. We are describing "average books" based on sufficient available data that helps us characterize them with a high level of confidence. In the shared print context, there are so many collections and groupings that it may be difficult to estimate the pertinent values for each possible subcollection—defined for example by language, subject, or circulation history. That being said, the model does indeed work on less typical cases, but one must determine the specific values to enter that pertain to that subcollection.

Adjustments could be made for collections that may have higher (or lower) inherent risk. Indeed, it is possible to apply the model on a title-by-title basis if enough were known about each numerical parameter, although the practicality and the value of doing so is dubious. Alternatively, the model facilitates what/if analysis, so ranges of possible values (of error and loss rates, for example) can be considered.

---

[*] See footnote above for actual estimates based on circulation from ReCAP's facility.

Meeting the Targets

The targets calculated by our model—a *minimum* of four to ten copies of each title preserved—place a high bar on shared print consortia. Currently most shared print programs seek to register a commitment to one last copy and permit individual members to make their own decisions whether or not additional copies are wanted to meet local demands. It is unlikely that individual shared print programs as currently configured will alone have the means or the administrative will to meet the standard for preservation of titles described here. This study emphasizes the need for shared print programs to further coordinate their efforts across multiple consortia in order to attain an appropriate number of commitments in a region, a country, or even worldwide.

Current and Future Research

A vexing problem in long-term efforts is the need to evaluate progress within a meaningful planning horizon. The community cannot realistically commit to a fifty-year waiting period before assessing whether the shared print enterprise has succeeded or not. Further research on risk factors is important to this.

The assumptions and numerical data behind each probability factor included in our model must be questioned, reevaluated, and revised: better data entered into our tool can give more refined results more quickly than waiting fifty years. In this paper, we have proposed methods for determining the minimum viable retention levels based on various levels of overall risk, in order to achieve a given level of confidence that a viable copy remains available at some point in the future. At this system-wide level, we consider only a few broad factors: storage conditions, starting age of materials, and estimated risk of loss.

As mentioned earlier, we used data from the EAST study to estimate probability of usability (degradation) curves, where we defined usability in a conservative manner. As more information becomes available about how books degrade over time in various storage conditions—including the present-day condition of books that have been kept in environmental conditions that are not well documented (and possibly not well controlled)–the curves that describe probability of usability should be refined and minimum retention requirements recalculated. In the near term, print archives research should be attentive to the trends in the findings from these studies. A library collection is a large, heterogenous gathering of papers that have been amassed and stored mostly in undocumented environmental conditions over hundreds of years. The technology to collect and analyze environmental data in useful form has been available only for a few decades. It is not realistic to expect a simple or universal answer to the question of how paper degrades. It is important to understand if the trends in research on this topic point towards overall better or worse outcomes compared to the current, limited data, especially where further research highlights subsets of materials for which there are signals for concern.

Other risks excluded from our model must also be revisited. For example, there is a notable risk from factors that cannot be anticipated or controlled, such as natural disaster or armed conflict, which have major and irreversible impacts on cooperative preservation efforts. To the degree that the library community can improve its knowledge of risk factors and develop controls, the outcomes of the shared print enterprise can be more predictable.

Some evaluation will be required after the first fifty-year milestone. In particular, the age to condition estimates need to be adjusted as items spend a greater percentage of their

time in good quality environments. The EAST data, and thus our usability curve, is based on data collected on materials that have been stored in the open stacks. Putting books into good quality environments reduces the rate of degradation for the period of time that they are in that environment. Once materials spend a significant portion of their lives in good quality environments, the degradation curve may need to be adjusted based on observed usability in the future. In the grand scheme, storage facilities are still new, and the impact of good quality environments is just beginning to adjust the usability estimates made here, but not enough to affect our generalizations. That will no longer be true after the first milestone. For example, a book that spends its first fifty years in the stacks and the next fifty in storage should appear closer to a 65-year-old book than a 100-year-old book. A book that is put into storage immediately may appear closer in condition to a book a quarter of its age that has spent its entire life in the stacks. No good quality storage facility has existed for fifty years yet, and many are younger than twenty years old.

The large impact of good quality environments also highlights the need in shared print management for a better way to determine what is in storage and what is in stacks. It is often unknown, copy by copy, which copies committed for shared print are in open library stacks and which are in storage facilities.

Shared print programs will also confront a fundamental not-enough-copies problem, the lack of sufficient copies of a work to meet preferences for risk from the outset. Solving the problem of what to do with titles that cannot reach the desired P1 will require strategies to mitigate or recover from information loss. Said another way, research on preservation strategies takes on a renewed importance in the shared print environment, so that we understand which methods and what resource levels are effective for addressing collections risk factors.

Gathering information on, or improving the forecasting of, risks can become an exercise with diminishing returns, however; libraries simply cannot fully predict or control future conditions. Consequently, development of meaningful interim targets and well-designed plans for validation of shared print archive holdings are important. Closely coupled to this is preservation science research that focuses on material properties of collection items in relation to their bibliographic identity, such as the Assessing the Physical Condition of the National Book Collection project coinvestigated by the Library of Congress and ReCAP (https://nationalbookcollection.org/overview). This effort connects the bibliographic focus of shared print to the material factors that determine preservation outcomes.

Finally, managers of libraries and shared print consortia need to review at the highest levels the costs and benefits of reducing risk. What is an acceptable level of loss? What will it cost to meet that threshold? Where libraries collectively hold ten or more copies of a title, will reducing the holdings still meet the current need for access? Where libraries hold fewer than ten copies of a title, what extra preservation measures, at what cost, are feasible to retain them? Ultimately this study is a tool in a broader range of decisions about the future of print collections.

Conclusions

Our research was motivated by the desire of the research library community to gain a better understanding of the number of copies of a monograph that need to be retained in a shared print arrangement to ensure a high probability of long-term availability and usability. Relying on the literature and our own studies, we identified four factors as critical for the long-term survival of monographs: (1) on-shelf probability; (2) bibliographic inaccuracy; (3) physical

loss or irreparable damage; and (4) gradual physical deterioration, which depends on the initial condition and ongoing storage environment. We incorporated these factors into a flexible decision support tool to help managers of shared print consortia develop targets for the number of copies of a monograph title they would need to retain in order to have a high level of confidence that at least one usable copy will remain at the selected time horizon. The tool is flexible, allowing decision-makers to input their own estimates of various risk factors, informed by available data.

We utilized the tool to perform calculations for a range of age and time horizon combinations, from a new book with a desired retention time of fifty years to a 100-year-old book with a desired retention time of 200 years (see appendix A). The results show that 10 copies would satisfy the minimum requirement or more to reach a 99.8% chance of survival of one usable copy in all of the modeled situations, even if all copies were held in open library stacks. Fewer copies may be needed if the selected time horizon is shortened, if the book is newer, or if at least some copies are stored in environmentally controlled storage rather than in open library stack conditions.

This research highlighted especially the large impact that closed-stack, environmentally controlled storage can have on the preservation of books. These conditions reduce the level of risk for our first three factors to a statistically insignificant level, and they reduce the rate of deterioration to a quarter of that for storage in typical open library stacks.

It was especially challenging to find reliable data characterizing the magnitude of each risk factor. Further research is needed both to test and verify the data used and to adjust the data as the implications of changes in the manufacture, storage, and use of monographs become evident.

Finally, and most importantly, this tool is only one facet of a much larger decision-making process confronting managers of libraries and shared print consortia. The model can calculate probabilities of survival, but managers must decide and agree on time horizons, tolerance for risks, and the cost/benefit trade-off of measures to retain titles into the future.

Acknowledgements

Lillian Dong made extensive contributions to the development of the spreadsheet tool. She was supported by research funds from the College of Engineering and the Haas School of Business at University of California, Berkeley.

Our research benefited greatly from thoughtful feedback and advice from many colleagues. We would like to thank especially those who participated on advisory panel discussions in the fall of 2021: Doug Brigham, Elise Calvi, Daniel Dollar, Dyani Feige, Terese Heidenwolf, Liz Hayden, Robert H Kieft, Jeff Kosokoff, Stephanie Lamson, Shari Laster, Lorrie McAllister, Sherri Michaels, Heather Parks, Todd Pattison, Peggy Seiden, Maggie Mason Smith, Steve Smith, Susan Stearns, Karla L. Strieb, Hannah Tashjian, Caitlin Tillman, Marie Waltz, Heather Weltin, and Alison Wohlers.

Thanks to Tom Clareson, who helped host early meetings, and Jennifer Hain Teper, Bobbie Pillette, and Katie Risseeuw, who contributed to an earlier group that did foundational work. Also, thanks to Robert Kieft and Oya Rieger, who worked with our group earlier in the project.

The Partnership for Shared Book Collections adopted, facilitated, and patiently encouraged this project.

Appendix A

The following table shows results from running the spreadsheet tool using the parameters described in this paper:

- Without item-by-item verification, the probability the copy can actually be located at the time of analysis is 97%
- Without item-by-item verification, the probability that bibliographic inaccuracy will result in selecting a copy for retention that is not the intended item is 0.1%
- The annual loss rate over the period if the copy is kept in stacks is 0.01% and 0.00% if kept in storage
- Deterioration over time progresses along a curve relative to the age of the copy according to the appropriate curve shown in Figure 7
- Deterioration is only 25% as large if the copy is kept in storage assuming that it is at about 20°F cooler than open library stacks

Within each time horizon, calculations are made for books that at the time of analysis are new, 50 years old, and 100 years old. Pink indicates that the probability (P1) of at least one usable copy remaining at the end of the designated time horizon (*T*) falls below 99.8%; blue that the probability exceeds 99.8%.

50 Year Horizon										
New Books										
	1 Copies	2 Copies	3 Copies	4 Copies	5 Copies	6 Copies	7 Copies	8 Copies	9 Copies	10 Copies
Stacks	82.77431%	97.03276%	99.48887%	99.91195%	99.98483%	99.99739%	99.99955%	99.99992%	99.99999%	100.00000%
Storage	97.30909%	99.92759%	99.99805%	99.99995%	100.00000%	100.00000%	100.00000%	100.00000%	100.00000%	100.00000%
50-Year-Old Books										
	1 Copies	2 Copies	3 Copies	4 Copies	5 Copies	6 Copies	7 Copies	8 Copies	9 Copies	10 Copies
Stacks	60.53314%	84.42367%	93.85251%	97.57378%	99.04245%	99.62208%	99.85085%	99.94113%	99.97677%	99.99083%
Storage	77.41734%	94.90023%	98.84834%	99.73992%	99.94127%	99.98674%	99.99700%	99.99932%	99.99985%	99.99997%
100-Year-Old Books										
	1 Copies	2 Copies	3 Copies	4 Copies	5 Copies	6 Copies	7 Copies	8 Copies	9 Copies	10 Copies
Stacks	49.47022%	74.46742%	87.09844%	93.48087%	96.70590%	98.33550%	99.15893%	99.57501%	99.78525%	99.89149%
Storage	60.30455%	84.24271%	93.74507%	97.51708%	99.01439%	99.60876%	99.84470%	99.93835%	99.97553%	99.99029%
100 Year Horizon										
New Books										
	1 Copies	2 Copies	3 Copies	4 Copies	5 Copies	6 Copies	7 Copies	8 Copies	9 Copies	10 Copies
Stacks	60.23121%	84.18443%	93.71034%	97.49868%	99.00525%	99.60440%	99.84268%	99.93743%	99.97512%	99.99010%
Storage	94.37419%	99.68350%	99.98219%	99.99900%	99.99994%	100.00000%	100.00000%	100.00000%	100.00000%	100.00000%
50-Year-Old Books										
	1 Copies	2 Copies	3 Copies	4 Copies	5 Copies	6 Copies	7 Copies	8 Copies	9 Copies	10 Copies
Stacks	49.22348%	74.21745%	86.90852%	93.35260%	96.62468%	98.28613%	99.12976%	99.55812%	99.77563%	99.88607%
Storage	72.85088%	92.62925%	97.99891%	99.45672%	99.85250%	99.95996%	99.98913%	99.99705%	99.99920%	99.99978%

100-Year-Old Books										
	1 Copies	2 Copies	3 Copies	4 Copies	5 Copies	6 Copies	7 Copies	8 Copies	9 Copies	10 Copies
Stacks	47.96938%	72.92814%	85.91434%	92.67114%	96.18675%	98.01594%	98.96768%	99.46288%	99.72053%	99.85459%
Storage	57.87940%	82.25855%	92.52720%	96.85241%	98.67422%	99.44157%	99.76479%	99.90093%	99.95827%	99.98242%
150 Year Horizon										
New Books										
	1 Copies	2 Copies	3 Copies	4 Copies	5 Copies	6 Copies	7 Copies	8 Copies	9 Copies	10 Copies
Stacks	48.97796%	73.96752%	86.71770%	93.22310%	96.54229%	98.23580%	99.09987%	99.54074%	99.76567%	99.88044%
Storage	91.01814%	99.19326%	99.92754%	99.99349%	99.99942%	99.99995%	100.00000%	100.00000%	100.00000%	100.00000%
50-Year-Old Books										
	1 Copies	2 Copies	3 Copies	4 Copies	5 Copies	6 Copies	7 Copies	8 Copies	9 Copies	10 Copies
Stacks	47.73012%	72.67859%	85.71913%	92.53541%	96.09827%	97.96057%	98.93399%	99.44280%	99.70875%	99.84776%
Storage	67.86361%	89.67252%	96.68112%	98.93343%	99.65724%	99.88985%	99.96460%	99.98862%	99.99634%	99.99883%
100-Year-Old Books										
	1 Copies	2 Copies	3 Copies	4 Copies	5 Copies	6 Copies	7 Copies	8 Copies	9 Copies	10 Copies
Stacks	47.73012%	72.67859%	85.71913%	92.53541%	96.09827%	97.96057%	98.93399%	99.44280%	99.70875%	99.84776%
Storage	55.30155%	80.02049%	91.06947%	96.00819%	98.21572%	99.20246%	99.64351%	99.84065%	99.92878%	99.96816%
200 Year Horizon										
New Books										
	1 Copies	2 Copies	3 Copies	4 Copies	5 Copies	6 Copies	7 Copies	8 Copies	9 Copies	10 Copies
Stacks	47.49205%	72.42915%	85.52311%	92.39848%	96.00860%	97.90420%	98.89954%	99.42217%	99.69659%	99.84069%
Storage	85.84795%	97.99719%	99.71656%	99.95989%	99.99432%	99.99920%	99.99989%	99.99998%	100.00000%	100.00000%
50-Year-Old Books										
	1 Copies	2 Copies	3 Copies	4 Copies	5 Copies	6 Copies	7 Copies	8 Copies	9 Copies	10 Copies
Stacks	47.49205%	72.42915%	85.52311%	92.39848%	96.00860%	97.90420%	98.89954%	99.42217%	99.69659%	99.84069%
Storage	62.78090%	86.14739%	94.84418%	98.08105%	99.28578%	99.73418%	99.90106%	99.96318%	99.98629%	99.99490%
100-Year-Old Books										
	1 Copies	2 Copies	3 Copies	4 Copies	5 Copies	6 Copies	7 Copies	8 Copies	9 Copies	10 Copies
Stacks	47.49205%	72.42915%	85.52311%	92.39848%	96.00860%	97.90420%	98.89954%	99.42217%	99.69659%	99.84069%
Storage	51.30719%	76.29010%	88.45498%	94.37841%	97.26269%	98.66713%	99.35099%	99.68398%	99.84612%	99.92507%

Notes

1. Candace Arai Yano, Zuo-Jun Max Shen, and Stephen Chan, "Optimising the Number of Copies and Storage Protocols for Print Preservation of Research Journals," *International Journal of Production Research* 51, no. 23–24 (November 18, 2013): 7456–69, https://doi.org/10.1080/00207543.2013.827810.

2. Lorcan Dempsey, Brian Lavoie, Constance Malpas, Lynn Silipigni Connaway, Roger C. Schonfeld, J. D. Shipengrover, and Günter Waibel, *Understanding the Collective Collection: Towards a System-Wide Perspective on Library Print Collections* (Dublin, Ohio: OCLC Programs and Research, 2013), 33, 99, http://www.oclc.org/research/publications/library/2013/2013-09r.html.

3. Zachary Maiorana, Ian Bogus, Mary Miller, Jacob Nadal, Katie Risseeuw, and Jennifer Hain Teper, "Everything Not Saved Will Be Lost: Preservation in the Age of Shared Print and Withdrawal Projects," *College & Research Libraries* 80, no. 7 (2019): 945–72, https://doi.org/10.5860/crl.80.7.945.

4. Ibid., 955.

5. Roger C. Schonfeld and Brian F. Lavoie, "Books without Boundaries: A Brief Tour of the System-Wide Print Book Collection," *The Journal of Electronic Publishing* 9, no. 2 (2006), https://doi.org/10.3998/3336451.0009.208.

6. Ian Bogus, Internal Analysis of Publication Dates in ReCAP, 2021.

7. Brian F. Lavoie, Lorcan Dempsey, and Constance Malpas, "Reflections on Collective Collections," *College & Research Libraries* 81, no. 6 (2020), https://doi.org/10.5860/crl.81.6.981), 988.

8. Schonfeld and Lavoie, "Books without Boundaries."

9. Lynn Silipigni Connaway, Edward T. O'Neill, and Chandra Prabha, "Last Copies: What's at Risk?*," *College & Research Libraries* 67, no. 4 (July 1, 2006): 370–79, https://doi.org/10.5860/crl.67.4.370.

10. Evan M. Anderson, "A Marking Heuristic for Materials in a Shared Print Agreement," *Library Resources & Technical Services* 61, no. 1 (2017), 5, https://doi.org/10.5860/lrts.61n1.04; Susan Stearns, Matthew Revitt, and Kirsten Leonard, "Taking Shared Print to the Next Level: The Partnership for Shared Book Collections," *Journal of Library Administration* 60, no. 7 (2020), https://doi.org/10.1080/01930826.2020.1803020; Helen N. Levenson, "Michigan Shared Print Initiative and Greenglass for Groups for Data Analysis in Developing a Collaborative Collective Collection," *Journal of Interlibrary Loan, Document Delivery & Electronic Reserve* 25, no. 3–5 (2015), 89-105, https://doi.org/10.1080/1072303x.2016.1254701; Rick Lugg, "Data-Driven Deselection for Monographs: A Rules-Based Approach to Weeding, Storage, and Shared Print Decisions," *Insights: The UKSG Journal* 25, no. 2 (2012)198–204, https://doi.org/10.1629/2048-7754.25.2.198.

11. Tony Horava, "What Is a 'Collection' Nowadays?," *Technicalities* 37, no. 6 (2017): 14–17.

12. Rick Lugg, "Remarkable Acceleration of Shared-Print," *OCLC Blog Next*, 1 March 2018, https://blog.oclc.org/next/the-remarkable-acceleration-of-shared-print/.

13. Michael Levine-Clark, "Access to Everything: Building the Future Academic Library Collection," *Portal: Libraries and the Academy* 14, no. 3 (2014), 426, https://doi.org/10.1353/pla.2014.0015.

14. Dempsey et al., *Understanding the Collective Collection,* 3.

15. Bernard F. Reilly. "Preserving America's Print Resources – Progress, Challenges and Necessary Measures in North America," *Bibliothek Forschung Und Praxis* 41, no. 3 (2017), https://doi.org/10.1515/bfp-2017-0043.

16. Paula Sullenger, "Is the Era of True Library Sharing within Reach?," *Technicalities* 37, no. 4 (2017): 1,4–6.

17. Yano, Shen, and Chan, "Optimising the Number of Copies and Storage Protocols for Print Preservation of Research Journals."

18. Canadian Conservation Institute, "Preventive Conservation and Risk Management: Agents of Deterioration" (Ottawa: Canadian Conservation Institute, 2022), https://www.canada.ca/en/conservation-institute/services/agents-deterioration.html.

19. Zachary Maiorana et al., "Everything Not Saved Will Be Lost: Preservation in the Age of Shared Print and Withdrawal Projects," *College & Research Libraries* 80, no. 7 (2019): 945–72, https://doi.org/10.5860/crl.80.7.945.; Cristina Duran-Casablancas, Josep Grau-Bové, and Matija Strlič, "Accumulation of Wear and Tear in Archival and Library Collections. Part I: Exploring the Concepts of Reliability and Epidemiology," *Heritage Science* 7, no. 1 (March 1, 2019): 10, https://doi.org/10.1186/s40494-019-0252-3.; José Luiz Pedersoli, Catherine Antomarchi, and Stefan Michalski, *A Guide to Risk Management of Cultural Heritage* (Ottawa: ICCROM; Government of Canada, Canadian Conservation Institute, 2016), https://ocm.iccrom.org/documents/guide-risk-management-cultural-heritage.; Agnes W. Brokerhof and Anna E. Bülow, "The QuiskScan—a Quick Risk Scan to Identify Value and Hazards in a Collection," *Journal of the Institute for Conservation* 39, no. 1 (2016): 18–28, https://doi.org/10.1080/19455224.2016.1152280.; Anna E. Bülow, "Collection Management Using Preservation Risk Assessment," *Journal of the Institute of Conservation* 33, no. 1 (2010): 65–78. ; and Stefan Michalski and José Luiz Pedersoli Jr, *The ABC Method: A Risk Management Approach to the Preservation of Cultural Heritage* (Ottawa: Canadian Conservation Institute

and ICCROM, 2016), https://www.iccrom.org/publication/abc-method-risk-management-approach-preservation-cultural-heritage.

20. Sara Amato and Susan Stearns, "Documenting the Stewardship of Libraries: The Eastern Academic Scholars' Trust Validation Sample Studies," *Collaborative Librarianship* 10, no. 3.4 (2018), https://digitalcommons.du.edu/collaborativelibrarianship/vol10/iss3/4.

21. Sherri Michaels and Becca Neel, "Conducting an Inventory with Shared Print in Mind," *Collection Management* 46, no. 2 (April 3, 2021), https://doi.org/10.1080/01462679.2020.1818343.

22. Ibid.

23. Ibid.

24. Sherri Michaels, Email to Ian Bogus, October 2, 2020.

25. Amato and Stearns, "Documenting the Stewardship of Libraries," 163.

26. "Assessing Book Condition," Validation Sample Study (Eastern Academic Scholars'," n.d.), https://sites.google.com/eastlibraries.org/validationsamplestudy/training-materials/assessing-book-condition.

27. Alex W. McKenzie, "Permanent Papers - From Then to When?" Australian Institute for the Conservation of Cultural Material (AICCM) *Bulletin* 16, no. 4 (1990): 25–32, https://doi.org/10.1179/bac.1990.16.4.003; Martin A. Hubbe, "Acidic and Alkaline Sizings for Printing, Writing, and Drawing Papers," *Book and Paper Group Annual* 23 (2004): 139–51, https://cool.culturalheritage.org/coolaic/sg/bpg/annual/v23/bpga23-24.pdf.

28. J. Tétreault, P.Bégin, S. Paris-Lacombe, and A.-L. Dupont, "Modelling Considerations for the Degradation of Cellulosic Paper," *Cellulose* 26, no. 3 (February 1, 2019): 2013–33, https://doi.org/10.1007/s10570-018-2156-x.

29. Stefan Michalski, "Double the Life for Each Five-Degree Drop, More than Double the Life for Each Halving of Relative Humidity," in *ICOM Committee for Conservation 13th Triennial Meeting Rio de Janeiro 20–27 September 2002* (London: James & James, 2002): 66–72, https://www.icom-cc-publications-online.org/2187/Double-the-life-for-each-five-degree-drop-more-than-double-the-life-for-each-halving-of-relative-humidity.

30. Adam B. Smith, "2021 U.S. Billion-Dollar Weather and Climate Disaster in Historical Context," in *Beyond the Data [Blog* (Washington DC: National Oceanic and Atmospheric Administration, 2022), https://www.climate.gov/news-features/blogs/beyond-data/2021-us-billion-dollar-weather-and-climate-disasters-historical.

31. Brian F Lavoie, Constance Malpas, J. D. Shipengrover, and OCLC Research, *Print Management at "Mega-Scale": A Regional Perspective on Print Book Collections in North America* (Dublin, Ohio: OCLC Research, 2012), 20–22, http://www.oclc.org/research/publications/library/2012/2012-05.pdf.

32. Zachary Maiorana et al., "Everything Not Saved," 953.

33. Jacob Nadal and Bob Kieft, "Curating Collective Collections: Silvaculture in the Stacks, or, Lessons from Another Conservation Movement," *Against the Grain* 27, no. 1 (2015); 70–71, https://doi.org/10.7771/2380-176x.7020.

34. Andres Stauffer and Bob Kieft, "Curating Collective Collections–A Forest for the Trees: A Response to Jacob Nadal's 'Silvaculture in the Stacks.,'" *Against the Grain* 27, no. 5 (2015), 83, https://doi.org/10.7771/2380-176x.7207.

35. Mike Garabedian and Bob Kieft, "Curating Collective Collections–Shared Print and the Book as Artifact," *Against the Grain* 28 (2016), https://doi.org/10.7771/2380-176X.7296; Mike Garabedian and Bob Kieft, "Curating Collective Collections–Shared Print and the Book as Artifact Part 2," *Against the Grain* 28 (2016): 73, https://doi.org/10.7771/2380-176x.7390.

36. Jennifer Hain Teper, "Considering 'Sameness' of Monographic Holdings in Shared Print Retention Decisions," *Library Resources & Technical Services* 63, no. 1 (2019): 29–45, https://doi.org/10.5860/lrts.63n1.29.

37. Paul Conway, "Preserving Imperfection: Assessing the Incidence of Digital Imaging Error in HathiTrust," *Preservation, Digital Technology & Culture* 42, no. 1 (2013) 17–30, https://doi.org/10.1515/pdtc-2013-0003.

38. Sara Gould and Richard Ebdon, eds., "IFLA/UNESCO Survey on Digitization and Preservation," in *International Preservation Issues 2*, 1999, 27, https://cdn.ifla.org/wp-content/uploads/files/assets/pac/ipi/ipi2%20vers2.pdf.

39. Thomas H. Teper and Vera S. Kuipers, "Exploring the Impact of Digitization on Print Usage," *Library Resources & Technical Services* 65, no. 2 (2021), 47, https://doi.org/10.5860/lrts.65n2.36-51.

40. Abhishek Nagaraj and Imke Reimers, "Digitization and the Demand for Physical Works: Evidence from the Google Books Project," *SSRN Electronic Journal*, 2019, 20, https://doi.org/10.2139/ssrn.3339524.

41. Schonfeld and Lavoie, "Books without Boundaries."

42. Chris Erickson, "Light, Dark and Dim Archives: What Are They?" *Digital Preservation Matters*, May 6, 2013, http://preservationmatters.blogspot.com/2013/05/light-dark-and-dim-archives-what-are.html.