OPTIMIZING THE NUMBER OF COPIES FOR PRINT PRESERVATION OF RESEARCH JOURNALS

Candace Arai Yano
IEOR Department and the Haas School of Business

Z.J. Max Shen
IEOR Department

Stephen Chan
IEOR Department

University of California
Berkeley, CA 94720

October 2008

OPTIMIZING THE NUMBER OF COPIES FOR PRINT PRESERVATION OF RESEARCH JOURNALS

Abstract

Academic libraries are reducing their holdings of print journals as more of this material becomes available electronically, but librarians, researchers and other users of this material, such an electronic archivists of journals, are concerned that some copies remain available. Journal archivists are especially concerned about preservation of "clean" copies that retain full information accuracy from the vantage point of the researcher.

We describe the results of research project designed to provide guidelines and insight to decision-makers in this context. As a prelude, we report briefly on statistical analysis of "defects" in the pages of 25 journals for their entire publication history. This provides a backdrop and motivation for our approach to the problem. We then present models for two storage protocols, both of which have the goal of minimizing the cost of ensuring, with a high probability, survival of at least one copy for a specified time horizon. One protocol involves archiving only clean copies in a secure environment, and the second protocol is a hybrid approach that combines clean copies with backup copies that can be "cleaned up" to replace clean copies that are lost or damaged. We also discuss other domains where our methodology can be applied.

Keywords: reliability, survival probability, inventory, research journals, print preservation, digital preservation

OPTIMIZING THE NUMBER OF COPIES FOR PRINT PRESERVATION OF RESEARCH JOURNALS

INTRODUCTION

Both librarians and researchers have been concerned about the preservation of archival research materials in print form. Although the amount of archival material that is available in electronic form is growing exponentially, availability of the original print artifact is important to various constituencies for various reasons. In some disciplinary fields, the original artifact may bear significant information that is not evident in the electronic copy, and in other disciplinary fields, access to good-quality renditions of graphical images may be critical. For a digital archivist that publishes the material in electronic form, the need for original print copy lies in the expectation that the material may need to be rescanned from time to time as technology changes.

Preservation of research materials in print form is taking on a new urgency. Atkinson (2001) warns of an impending crisis in libraries' ability to preserve paper-based materials. The disappearance of these collections would be a big loss for researchers. Improper management of the library materials has already led to the disappearance of a significant number of distinct titles. In a study on book deterioration and loss, O'Neill and Boomgaarden (1995) sampled 1,935 books at libraries in Ohio published between 1851 and 1939 and representing 872 distinct titles. They found that 12 percent of the books were unavailable for physical examination because they were lost, missing, or weeded from their collections.

In recent years, many libraries have deaccessioned second copies of journals, and some have started to deaccession even their first copies of journals that are available

electronically under the assumption that researchers no longer need the print copy because they can access the electronic copy instead. Although some large libraries will maintain extensive paper collections for the foreseeable future, the cost of doing so at the local level will eventually become too expensive. It is unrealistic for individual libraries to devote scarce resources to preserve massive and separate holdings of printed materials. As a result, university libraries worldwide are starting to form consortia to ensure retention of the "last copy" of various materials. The University of California's Northern and Southern Regional Library Facilities, built in the early 1980s, represent the first major examples of facilities shared among multiple campuses. In recent years, there has been a growing trend toward building shared facilities as opposed to individual facilities.

Broadly speaking, there are two types of storage arrangements being used by consortia: depositories and repositories. A depository is a shared or cooperative storage facility where the depositing libraries retain ownership of the materials. A repository, sometimes called a last-copy repository, is a facility serving a regional area or a group of participating libraries where ownership of the materials transfers to the repository (Payne 2005).

Several major shared depositories exist. For example, the Washington Research Library Consortium (WRLC) operates a digital library system and a shared storage facility for eight university libraries in Washington, DC. The Research Collections and Preservation Consortium (ReCAP) facility is jointly owned by the New York Public Library, Columbia University, and Princeton University, and is located at and operated by Princeton University. The Minnesota Library Access Center (MLAC) operated by Minitex at the University of Minnesota, includes volumes from public libraries as well as

academic materials. NELINET, a cooperative of more than 600 academic, public, and special libraries in the six New England states, manages the New England Regional Depository. See Payne (2004) for more discussion of different types of shared depositories.

Compared with shared depositories, the number of repositories is limited. In the case of several of these repositories (e.g., the University of California's Northern and Southern Regional Library Facilities), although the materials had been spread across multiple campuses, they were, in principle, already owned by a single entity before they were placed into shared storage. Other repositories are truly inter-institutional in nature. The Center for Research Libraries is a consortium of over 230 university, college and independent research libraries that acquires and archives journals, newspapers, infrequently-accessed books and other documents in both paper and electronic form. A repository in Hong Kong (Electronic Resource Academic Library Link) that emphasizes e-books provides a central storage facility for several universities, both public and private. This consortium of libraries makes joint purchases of shared materials. The Five Colleges, Inc., repository (http://www.fivecolleges.edu/sites/depository/) was established by five institutions of higher education in western Massachusetts as a shared facility for lesser-used materials.

The idea of establishing a distributed network of print archives has been gathering momentum recently. Grants have been given to librarians to study the cooperative selection and deselection of science serials (Roberts 1988 and Chrzastowski et al. 2007), and there have been studies on the feasibility and effectiveness of collaborative collection development (Hightower and Soete 1995, Seiden et al. 2002, Seaman 2005, and Holley

2003). The library community recognizes that cooperative arrangements, rather than individual, redundant local endeavors, represent the only pragmatic strategy in the future. Reilly (2002) suggests using the national energy grid as a conceptual model for forming of inter-institutional consortia network. He argues that the energy industry and consortia libraries are similar since they both deal with issues of access, conservation, and delivery of resources on a national basis. The energy industry as a whole does a very effective job of balancing individual self-interest against public responsibility. Reilly believes that the key for such success is the national energy grid. The grid consists of the overall network for the supply and delivery of energy resources for the nation: the sum of thousands of (mostly) privately owned systems of cables, conduits, pipelines, storage facilities.

Our work was motivated by JSTOR's concern about identifying a cost-effective strategy to secure and protect a sufficient number of print copies of a journal or collection of journals so that, with very high probability, all of the printed material within the journals will be available for an appropriate time into the future. At the time we started this study, JSTOR already owned or controlled two sets of the journals for which it provides electronic access, but JSTOR management was seeking to understand whether this was enough, and if not, then how many additional copies would be needed.

In this paper, we focus on preservation of archival research journals, but all of the analytical methodologies can be readily adapted to other types of archival materials. In the case of digital materials, defects may be introduced at a higher rate than in print materials (Rosenthal, 2008) and additional complications arise in the transition from one electronic format to another (Sivathanu et al. 2005), but the analytical methodologies are still useful in modeling many aspects of the problem.

Our research is based on the premise that an organization or consortium would like to identify a cost-effective strategy for ensuring that it has secured and protected a sufficient number of print copies of a journal so that, with a very high probability, all of the printed material within the journal will be available for a specified time horizon. We take the viewpoint of an individual or organization that is concerned not only about the availability of the materials but also its informational accuracy. Thus, we are concerned about availability of "perfect copy," i.e., copy that is as good as new from the vantage point of the information that it carries for the researcher.

We develop a quantitative modeling framework and analysis methods for addressing this problem. The framework accounts for the fact that extant copies of journals are unlikely to be in flawless condition, and that both environmental and catastrophic risks must be considered—along with economic factors—in making these decisions.

In some respects, our problem may look similar to the problem of spare parts acquisition management at the end of the product life cycle. For example, after the production of a complex product such as a helicopter has ceased, the manufacturer is expected to provide (usually at a cost) replacements for failed components. It is generally much less expensive to produce these spare parts while the product is still in production, and for this reason, many manufacturers stockpile spare parts at this time. How many units of each component should be stockpiled? This decision bears some similarity to the decisions faced by a digital archivist or library consortium. However, the spare parts scenario differs significantly from the journal preservation scenario. For example, demand for spare parts is generated by failures in the field that do not

necessarily have any relation to the number of spare parts in inventory, whereas in the journal preservation scenario, depletion of "spare" journals is caused by loss/damage of the spares themselves, and therefore depends upon the number of surviving copies. As another example, in the spare parts scenario, research articles almost always assume that the spare parts remain in perfect condition, but this is not necessarily true in the journal scenario. Thus, we cannot simply borrow solution techniques from the spare parts inventory management literature.

Recently we have seen a great deal of research on supply chain disruption management (e.g., Chopra and Sodhi 2004, Kleindorfer and Saad 2005, Tang 2005, Snyder and Shen 2007). However, the focus of this line of research is on issues of risk in supply chains and methods for mitigating these risks, and no attention has been given to situations where the stored items are irreplaceable. A very recent paper by Ruiz-Torres and Mahmoodu (2007) addresses the problem of optimizing the number of suppliers (analogous to journal copies) when the suppliers are unreliable but statistically independent.

There is also a link between the models that we plan to develop and models in the reliability literature concerning the optimal degree of redundancy. Inspired by reliability issues for Redundant Arrays of Inexpensive Disks (RAID), Baker et al. (2006) propose another reliability model for long-term storage failures that addresses a wide range of faults. They also provide a review of related literature on correlated and latent faults that arise when large volumes of data must remain unaltered yet accessible with low latency at low cost. Different from this line of research, our main technical complication here is that if any imperfect copies are utilized as a means of preserving information, their

defects are likely to be correlated. In particular, we expect that in comparison to "unpopular" articles, "popular" articles and the issues and volumes in which they are contained will have higher defect rates. Thus, the locations of defects are likely to be correlated from one copy of a journal to another. There is very limited research on optimal degrees of redundancy in systems with parallel resources when the reliability levels of the resources are correlated. For example, a recent proceedings paper by Sarper and Chacon (2006) treats the case of (only) two systems operating in parallel.

The remainder of this paper is organized as follows. The next section contains some preliminaries. We then provide a formal statement of the problem and related analysis. The paper concludes with a discussion of implementation and conclusions.

PRELIMINARIES

In this section, we present background information and an overview of statistical analysis that provides a backdrop and motivation for our approach to the problem. (Further details are available from the authors.)

Statistical Analysis of Defects in Journals

Before we began to develop our analytical models, we considered it prudent to analyze the condition of "off-the-shelf" journals. Among other things, we were interested in understanding the following:

- Frequency and types of defects as a function of journal age and probable usage rates.
- Correlation of defects across multiple copies of the same material.

We were able to obtain records of defects on a page-by-page basis for about 25 JSTOR journals. The defects had been recorded at the time the material was being

prepared for scanning. In virtually all cases, data were available for the entire publication history of the journal. Defects ranged from missing or torn pages at the more severe end of the spectrum to marks in the margin and blurred text at the other end of the spectrum. Except in the case of two journals (discussed in more detail later), the defect rate (defects per page) considering all defects was quite low, generally on the order of one defect per 10,000 to 100,000 pages. These statistics may not be reflective of journals in general, as JSTOR sought out relatively clean copies where possible, and some of the materials consisted of new issues or volumes obtained from publishers. However, a significant portion of the material was obtained from the libraries of major research universities where the usage of these materials was likely higher than at smaller universities or those with less emphasis on research. Consequently, our analysis gives us reason to believe that "off-the-shelf" journals will generally be in very good condition, but not in perfect condition.

As mentioned earlier, two journals were not in good condition and exhibited defect rates that were significantly higher than those of the other journals. One was a nursing journal that might be classified as a professional magazine. This publication contained significant amounts of advertising material, etc., and may have been treated more like a magazine than like a research journal by its users. The other journal was a medical journal. Its relatively poor condition may have been due to extensive use, but we did not have enough information to ascertain probable causes. It is possible that medical-related journals experience much more extensive use than journals in other disciplines, but we did not have access to defect data for any additional medical-related journals to confirm or refute this conjecture.

To assess whether defect rates were correlated with age and/or usage, we developed linear regression models. Here, we limited our study to the journals that were in good condition, as they appeared to be representative. For about 40% of the journals, the coefficient reflecting the increase in the defect rate with age was statistically significant at the $p = 0.05$ level. However, the absolute values of these coefficients were quite small. For example, a typical value of the coefficient would indicate that the defect rate would increase by 1% over a period of 1000 or more years. We regarded these numbers as being so small that they were not practically meaningful. Furthermore, for many journals, the coefficient was negative, indicating that the condition of older issues was *better* than that of newer issues. Thus, any connection between the defect rate and age of the journal could be construed as spurious.

Libraries do not have very accurate records of journal usage over the past decades or centuries. As a proxy for usage, we utilized total (worldwide) viewings (on JSTOR) of articles within each issue during a 24 month period. (Downloads were highly correlated with viewings, so it was sufficient to utilize viewings. Furthermore, a viewing would be the closest analog to physically retrieving a copy of the journal issue and looking at it.) The regression analyses provided no statistically significant relationships, probably because our proxy for usage was not reflective of actual usage over the lifetime of a journal issue.

Based on the above results and the reasonable assumptions that (i) the availability of the material in electronic form will lead to very low (physical) usage rates (and thus also low rates of quality degradation) in the future, (ii) the ongoing and significant movement of print materials from open stacks to less-accessible remote storage locations

will lead to a decline in damage and loss rates, and (iii) modern storage conditions (especially air conditioning and humidity control) should lead to much less deterioration due to environmental conditions than would have occurred in past generations, we do not specifically model degradation due to age or usage in detail, but simply include it as a (small) part of an aggregate loss rate from all risk factors.

We were, unfortunately, unable to obtain defect data on multiple copies of the same journal to test for possible correlation in the location of defects. Anecdotal evidence suggests that high-visibility articles and distinctive graphical material are more often cut or torn from journals than less visible and more mundane material. It is largely for this reason, and because we observed two journals (the medical-related journals) with significantly higher defect rates, that we consider only archiving strategies in which at least one "clean" copy (quality-checked and repaired, if necessary, on a page by page basis) is retained.

Estimates of Loss Rates

Accurate data on loss rates for library materials are not readily available. We contacted several major university libraries in an attempt to obtain data on loss rates, but were unable to collect enough reliable information to make useful estimates. One major university offered that it used an annual loss rate of 1% for planning purposes, although its observed loss rates appeared to be far less. Part of the challenge here is that although research libraries do have some data on known losses, extensive audits are rarely performed to determine actual losses. Moreover, pages within individual volumes are virtually never checked to ascertain their condition.

In view of these difficulties, we contacted an insurance company that is a major provider of policies to university libraries to obtain rate information from which we could extrapolate estimates of actual losses. Quoted (annual) rates in the neighborhood of $1200 per million dollars of valuation suggest insurance payouts in the neighborhood of $1000 per million, or loss rates of 0.001 annually. However, these payouts do not include losses covered by deductibles and losses for which no claims are filed (because the organization is unaware of losses or does not find it economical to file a claim). Consequently, we regard an annual loss rate of 0.001 to be the lower limit of the loss rates that would be achievable in practice. Even for locked-up materials, there are risks due to fire, plumbing problems, natural disasters, etc., which, in the aggregate, might lead to loss rates of approximately 0.001 annually.

We regard a 1% loss rate for circulating material to be a conservative (high) estimate, and use this number in some of our calculations with the recognition that its use will lead to conservative (high) numbers of copies to be archived.

PROBLEM STATEMENT AND ANALYSIS

Our definition of the problem evolved over time as we gathered and analyzed data, and as we learned more about the relative costs of maintaining material at different levels of security and access. Ultimately, we developed models for two fundamentally different protocols. We discuss each in turn.

Storing Multiple Perfect Copies

In our first model, we assume a strategy in which the digital archivist relies solely on multiple perfect (fully verified) copies, with the number to be determined. By

"verified," we mean that each copy has been quality-checked on a page-by-page basis. We also assume that any small defects are repaired as they occur, so the copies are maintained in "perfect" condition. Thus, the losses under consideration are actual physical losses or serious damage that would, in view of practical or economic considerations, render a volume irreparable. Each copy would be stored in a distinct location so as to minimize the effects of shared risks. Here, we do not specify the details of the storage conditions; we only require an estimate of the annual loss/damage rate for the selected storage conditions. Of course, the storage conditions and degree of security should be chosen so that the maintenance of "perfect" copy is manageable. The digital archivist seeks to find the minimum number of copies needed to ensure survival of at least one copy with a specified probability, α, and for a specified time horizon, T.

Both here and in the protocol described later in this section, we take the perspective of the journal archivist that is concerned about protection of information and not about the intactness of a particular journal in its entirety. Virtually all of the journals will be stored in the form of bound volumes, so we take a bound volume, which would typically include one year (or a portion thereof) of published material in a journal as the unit of analysis. The goal is to ensure (with a high probability) the survival of the information contents of each bound volume *independently*. One can specify higher survival probabilities for journals that are considered more important. But it may be impractical to ensure with a very high probability that all (say) 100 volumes of a long-standing journal survive for an extended time horizon. This would require even more copies than what we derive in our analysis here. We return to this point in the concluding section.

For simplicity, we assume the same annual probability, p, that any given material (e.g., a volume) would become lost, irreparably damaged, or otherwise unusable as a perfect copy. Under the assumptions of this model, if we start with N total copies, we can express the probability of having at least one copy survive for T periods analytically. The probability that a single copy survives for T periods is $(1-p)^T$, so the probability that it does not survive is $1-(1-p)^T$. The survival (or non-survival) of each individual copy is assumed to be independent of the others, so the probability that all N copies do not survive T periods is the product of individual probabilities of non-survival, i.e., $[1-(1-p)^T]^N$. The probability that at least one copy survives for T years is thus:

$$\text{Prob(at least one surviving copy after } T \text{ years)} = 1 - [1-(1-p)^T]^N$$

Recall that α is the target survival probability. Thus, to find the minimum number of copies required, we need to find the smallest value of N such that

$$1 - [1-(1-p)^T]^N > \alpha$$

which is equivalent to

$$1 - \alpha > [1-(1-p)^T]^N$$

Taking the natural log of both sides, we can determine number of required copies as the smallest value of N such that

$$\ln(1 - \alpha) / \ln[1-(1-p)^T] > N,$$

so we can express N as $\lceil \ln(1 - \alpha) / \ln[1-(1-p)^T] \rceil$ where $\lceil x \rceil$ is the smallest integer greater than or equal to x.

Table 1 shows a few example calculations for various combinations of time horizons, survival probabilities and annual loss rates.

Table 1: Number of Copies Required to Meet Target Survival Probability for the Given Time Horizon and Annual Loss Rate

	T = 50 years	100 years	100 years	200 years
annual loss rate	Survival Prob. 0.999999	Survival Prob. 0.999	Survival Prob. 0.999999	Survival Prob. 0.999999
0.001	5	3	6	8
0.005	10	8	15	31
0.010	15	16	31	96

These numbers are surprisingly large and the requirements implied by them are onerous. For example, if we desire a survival probability of 0.999999 for 100 years and copies of a journal have an annual loss rate of 0.005, we would need to retain at least 15 copies of the journal. For more esoteric journals, the question of feasibility arises: Are there even 15 complete copies of the journal in existence? Even for more widely distributed journals, the cost of verifying 15 copies on a page-by-page basis would be quite significant, and this process would have to be repeated for every journal in the relevant collection. We therefore need a method to reduce the required number of perfect copies while achieving the specified survival probability.

One way to limit the required number of perfect copies is to reduce the annual loss rate by increasing security. After we discussed this possibility with various librarians, it became clear that intermediate levels of protection, such as those utilized for special collections, would be extraordinarily expensive. Special collections, for which access is limited, require considerable labor for interacting with prospective borrowers, for retrieving materials from library stacks, and for monitoring the borrower if only in-library use is allowed. Although intermediate levels of security are expensive, "dark archives" in which copies are (essentially) not allowed to circulate at all, is even less

expensive than storage of normal circulating copies. There are several factors contributing to the low cost, including the possibility of using high-density storage (which saves considerable space), the possibility of locating the storage facility in a remote (lower-cost) location, and the minimal labor required to maintain the collection. Because intermediate levels of security are not economical but low and high security are relatively less expensive, this suggests that we take a mixed strategy approach.

Hybrid System

We consider a hybrid system as an alternative to the system with multiple perfect copies. The hybrid system consists of a combination of highly secured "locked-up" copies and less secured, possibly circulating, copies, The "locked-up" copies are verified (quality-checked and repaired, if necessary) on a page-by-page basis and maintained in "perfect" condition. The less secured "backup" copies are verified only at the issue level, not at the page level, to ensure that each journal series is complete at the outset. These less-secured copies serve as a backup for the locked-up copies: if and when a locked-up copy becomes lost or irreparably damaged, replacement material is pulled from the set of backup copies, fully verified, and then added to the locked-up set. Under this arrangement, until the bitter end, there would always be a verified, locked-up copy, which, when combined with available electronic copies, could be used as the basis for repairing the replacement material, as needed.

We recognize that the term "locked-up" is not standard terminology in the library community, but we use this term to communicate a key assumption of our model that not only are these copies fully verified at the outset, but also the level of security is such that these materials can be kept in perfect condition, and if any are lost or damaged, they will

be replaced with fully-verified material (essentially) immediately. Thus, one possible method of storage is what is called "dark archives" in the library community—an archive with no access to users. One could instead utilize "dim archives" with very restricted access, but in this case, it would be necessary to page-verify any borrowed material upon return, and to repair it if necessary, to ensure that it remains in perfect condition. Thus, "very dim archives" with extremely restricted access may be an economically viable option when labor costs are considered. However, moderately restrictive access policies would be very expensive due to the day-to-day operating costs mentioned earlier, as well as the labor required to keep the locked-up copies in perfect condition (i.e., ongoing page-level verification and repair).

Storage options for the backup copies include open shelves or remote storage without special access restrictions. These storage arrangements would have higher damage/loss rates than the storage for locked-up copies. Nevertheless, in view of the very low usage rates for print copies of journals when electronic copy is available, we do not expect noticeable micro-level deterioration or damage to the backup copies provided that they are stored under reasonable environmental conditions (e.g., air conditioning in humid regions) and with the usual library security provisions.

As in the system with multiple perfect copies, we do not specify the exact storage conditions for the locked-up or backup copies. For the locked-up copies, we only require relatively high security (and thus very low damage/loss rates) and a protocol by which the material is maintained in perfect condition. For the backup copies, our model requires that these materials be issue-verified at the outset, but there is no requirement or expectation that these journal series will remain complete; the model accounts for

damage and loss that would be expected to take place over time, which is reflected in the annual loss probability.

This hybrid system has several advantages. Locked-up copies are assumed to be (essentially) non-circulating so the annual loss rate for these copies is significantly lower than for circulating copies. This yields two benefits. First, relatively few locked-up copies are needed. Second, once the locked-up sets are established, relatively little labor is needed to secure them, unlike special collections. Likewise, backup copies do not need much special effort, apart from ensuring completeness at the outset. Another advantage is that the material does not need to be fully verified unless it is needed for the locked-up set.

The digital archivist seeks to find the Pareto frontier of locked-up and backup copies needed to ensure the survival of at least one copy (again, for each volume independently) with a specified probability, α, and for a specified time horizon, T. The Pareto frontier is the set of all combinations of number of locked-up copies (n_{locked}) and number of backup copies (n_{backup}) such that it is not possible to reduce either n_{locked} or n_{backup} without violating the target survival probability requirement. Our reason for approaching the problem in this way is that the costs for obtaining or otherwise gaining access rights to copies of journals are difficult to estimate. However, it is clear that locked-up copies will require greater effort and incur greater expense than backup copies. Furthermore, the cost tradeoffs vary from journal title to journal title. Thus, the best combination of n_{locked} and n_{backup} for one journal may be quite different than it is for another journal. The availability of the Pareto frontier allows the journal archivist to choose the best point on the curve for each journal. It is also important to point out that

the appropriate survival probability and time horizon may differ from journal to journal, and the journal achivist may choose those parameters consistently with its objectives.

For the hybrid model, we assume the annual loss probabilities for locked-up (p_{locked}) and backup copies (p_{backup}) differ, with $p_{locked} < p_{backup}$. In order to derive the Pareto frontier, we need to calculate the survival probabilities for various values of n_{locked} and n_{backup}.

To calculate the probability of survival for a given hybrid system (T, α, n_{locked}, p_{locked}, n_{backup}, p_{backup}) we enumerate the possible states of the hybrid system in each period, where the state of the system is defined by the pair ($n_{locked}(t)$, $n_{backup}(t)$), with t denoting the time period. In each period and for each state, we determine the probability that the system will be each of the possible states. Initially, at time 0, the hybrid system is in one state ($n_{locked}(0)$, $n_{backup}(0)$). After one period, the hybrid system will make a transition and be in one of $n_{locked}(0) + n_{backup}(0) + 1$ states. For example, if initially $n_{locked} = 2$ and $n_{backup} = 2$, then after one period the system can be in one of the following states: (0, 0), (1, 0), (2, 0), (2, 1), and (2, 2). We have to calculate the probability of the system making each possible transition, where the nature of the transition depends on how many "locked-up" copies and how many backup copies are lost in that period. For example, the system transitions from state (2,2) at time 0 to state (2,1) at time 1 if either (i) one backup copy is lost/damaged or (ii) one locked-up copy is lost/damaged and replaced by a backup copy. The calculations are perfomed starting at time 0, then progressing to period 1, then to period 2, and so forth until period T. To calculate the survival probability at time T, we add the probabilities associated with all

states in which at least one (verified) copy has survived. (This is equivalent to the sum of probabilities of all states except the (0,0) state.) Due to the enumerative nature of these calculations, it is not possible to express the survival probabilities using a formula. However, the calculations themselves are straightforward, albeit tedious.

Table 2 shows examples of Pareto frontiers for various values of p_{locked} and p_{backup} for a survival probability of 0.999999 and a time horizon of $T = 100$. Note that a minimum of two verified copies are needed (unless attrition makes this impossible). If and when one of the verified copies is lost, the other verified copy can be used as the basis for constructing a new verified copy from a copy drawn from the backups. Without a scheme in which two verified copies are maintained (until attrition makes this impossible), it is difficult to guarantee that a backup copy can be verified and repaired, unless a perfect copy of a "born electronically" file is available for the material. When such perfect, "born electronically" files are available (e.g., from the publisher), it may be possible to consider solutions in which only one verified print copy is maintained, assuming that a "perfect" version of the electronic copy can be maintained.

The numbers in the table illustrate the effects of p_{locked} and p_{backup} on the numbers of required copies. For example, doubling the loss rate of the backup copies (i.e., moving from the left column to the middle column) results in a modest increase in the numbers of backup copies needed for a given number of locked-up copies, and increasing the loss rate of locked-up copies (moving from the middle column to the right column) not only increases the number of locked-up copies that are needed but also increases the number or backup copies required. The difference between the second and third column also demonstrate the importance of keeping p_{locked} low in order to keep the

Table 2: Combinations of Locked-up and Backup Copies for Survival Probability = 0.999999 and Time Horizon T = 100

$p_{locked} = 0.001$ $p_{backup} = 0.005$		$p_{locked} = 0.001$ $p_{backup} = 0.01$		$p_{locked} = 0.005$ $p_{backup} = 0.01$	
n_{locked}	n_{backup}	n_{locked}	n_{backup}	n_{locked}	n_{backup}
2	8	2	13	2	22
3	5	3	8	3	19
4	3	4	4	4	17
5	2	5	2	5	15
6	0	6	0	6	13
				7	11
				8	9
				9	8
				10	6
				11	5
				12	4
				13	2
				14	1
				15	0

total number of archived copies within reasonable limits. Due to the probabilistic nature of the system, the changes are not proportional: a doubling or halving of a loss rate does not lead to a doubling or halving of the number of required copies. The changes may be more than or less than proportional to the changes in the loss rates.

To illustrate the practical benefits of the hybrid system, consider the example that we used for the model with multiple perfect copies, and assume further that the annual loss rate for backup copies is 0.01, so the right column in Table 2 is relevant. Whereas 15 locked-up copies were needed in the absence of backup copies, one could instead choose 7 locked-up copies along with 11 backup copies. This solution requires only 3 additional copies in total, but provides a 50+% reduction in the number of locked-up copies required. This significantly reduces the amount of page-level verification that

must be performed up front, and leaves many more copies in circulation for library patrons to use.

We note that finding the Pareto frontier can be done very quickly (usually in a matter of seconds) using a spreadsheet and because it identifies all undominated feasible solutions, it is straightforward to find the optimal solution for any arbitrary cost function. Furthermore, knowledge of the Pareto frontier can help the decision-maker identify a small set of configurations for which costs need to be estimated, and in some cases, even very rough estimates of costs are sufficient to identify the best one or two points on the Pareto frontier. Using a general integer program to solve the problem for a given cost function would likely take much more computing time and require much more sophisticated software. Thus, our methodology is very accessible to members of the academic library community who typically have no knowledge of probability theory or optimization.

IMPLEMENTATION AND CONCLUSIONS

Decision-making regarding archiving of research journals for the purpose of preservation will be an ongoing process, so we will never be able to claim that our analysis was implemented. However, our analysis has already strongly influenced the exploration of appropriate, economical, strategies by JSTOR as well as decision-makers for consortia organized for the purpose of preserving documents in digital form. In particular, our analysis shows clear advantages of using the proposed hybrid system for print archives, and this has helped decision-makers understand the need for, and value of, placing a small number of copies in "dark archives."

Earlier, we mentioned that a larger number of copies would be needed to maintain intactness of an entire journal series. To illustrate this point, suppose that we set α=0.999999 for each volume independently. Then, under the assumption that losses of the individual volumes are independent, the probability that all 100 volumes in a hypothetical journal survive is only α^{100}= 0.9999. This may be a perfectly acceptable survival probability. On the other hand, if one wishes to ensure with probability 0.999999 that all 100 volumes survive, then it would be necessary to archive 20 (rather than 15) perfect copies. The difference here is not dramatic because the initial target survival probability was already quite high, but in other instances, the differences may be much larger. Furthermore, each incremental copy becomes more difficult to secure, so even the five additional copies in this example could be quite problematic.

Our analysis is also useful in quantifying the need for multiple (often many) digital copies of documents because there is evidence that digital documents degrade more rapidly than print material (Rosenthal 2008), so the estimated loss rates that would need to be utilized in our analysis are higher than that for print materials.

Much more needs to be done to obtain better estimates of loss rates, but in the interim, our model can provide librarians and library consortia initial guidelines for planning before rapid de-accessioning of journals makes it impossible to provide the desired level of protection.

From a technical standpoint, it would be useful to generalize the model to consider catastrophic risks such as floods or fires that may affect multiple volumes at the same time. Events of this type are rare and typically affect only a small portion of an archive. Nevertheless, it would be valuable to understand the circumstances in which

consideration of correlation due to common-cause risks would significantly affect the decision of the number of copies to archive.

As part of this project, we developed spreadsheets that (1) compute survival probabilities for the model with multiple perfect copies; (2) compute survival probabilities for the hybrid model; and (3) determine the Pareto frontier for the hybrid model. These are available at (http://ieor.berkeley.edu/~shen/SurvivalPerfectCopies.xls and http://ieor.berkeley.edu/~shen/HybridSystemAnalysis.xls) or from the authors.

In concluding this paper, we discuss a few other potential applications of our hybrid model and broader lessons learned from our study. The framework of our hybrid model can be utilized—with some generalization—in other situations where preservation is the key goal. For example, there is increasing concern about the preservation of endangered animal and plant species whose populations have declined due to global warming and other environmental conditions. In the case of animals, the lock-up location might be a zoo, where the animals would be (essentially) protected from predators, and the "backups" might be held in a wildlife preserve, where they would be protected from poachers and hunters,but not from natural predators. (Of course, the animals may reproduce, which would add another element of uncertainty and complexity to the analysis.) Issues of this type have also been considered with the goal of maximizing biodiversity within a budget constraint by Weitzman (1998), who also discusses application of his model to library collections. Weitzman's model focuses on a snapshot at one point in time and does not consider long-term preservation issues.

The Norwegian government has built a "Doomsday Vault" in a mountainside on a remote island near the North Pole to preserve seeds of plant species (see

http://news.nationalgeographic.com/news/2008/02/photogalleries/seedvault-pictures/). Here, the seeds are carefully protected, yet will deteriorate over time, and they are stored in a single location. Questions arise as to how many seeds should be stored, when plants should be grown to generate new seeds, and whether the seeds should be in geographically dispersed locations to reduce shared risks. Some commercial seed companies are also making investments in seed preservation, but not nearly with the same level of security as the "Doomsday Vault." An adaptation of our hybrid model may be useful in studying how a dispersed set of organizations, most of which do not have the level of protection and security of the "Doomsday Vault," might collectively maintain plant seeds for posterity.

Another related example, although a bit more general, pertains to roster management of a sports team. This example was motivated by a situation in which all of the quarterbacks on a major college football team became injured, and a player who had played quarterback in high school but not in college was forced into the role of quarterback. A football team typically maintains a certain number of players for each position, and some of them are regarded as "reserves" because they would only be called in when the more skilled "active" players are injured. (They may also be called in when their team is winning by many touchdowns, but we ignore this possibility here.) Suppose that we have only these two categories of players and we are only modeling season-ending injuries. Coaches need to identify the right numbers of active players so that they can dedicate enough time and energy to coach these players hoping to maximize their performance. However, knowing that some of these players may get hurt during the season, the team should also identify enough reserve players (who receive less attention

and get less serious practice), so that they can be "upgraded" to active status when needed, thereby ensuring that at least one minimally qualified person is available for each key position.

There are also applications in the realm of maintaining digital files, far too numerous to mention here.

Finally, we turn to lessons learned from our study. To understand these lessons, it is useful to understand our mindset and experiences as we pursued the project. When we began our study, what we had in mind was a formal cost optimization model to find the minimum cost configuration to achieve JSTOR's target survival probability. Initially, we had no idea about which risk factors were dominant and significant, nor did JSTOR management and staff. So we had to perform our own, mostly unguided, investigation of various sources of risk, including the rather detailed statistical analysis described earlier in this paper, an exploration of the causes of theft from libraries, etc., to determine which risks to model and how to model them. We ultimately determined that deterioration of the printed copy itself, even if the material was in circulation, would be only a very minor component of the overall risk, and that probabilities associated with other risks such as mold that can develop in humid regions due to power losses, broken water pipes, fires, floods, etc. , would be extremely difficult to assess. This eventually led us to our decision simply to model risks in an aggregate manner.

The next challenge that we faced was estimating and modeling costs. Initially, we wanted to understand how adding additional security or improved library infrastructure would reduce risk and we explored these issues for some time. However, it soon became clear that with the number of journal copies that would need to be archived, it would be

financially infeasible for JSTOR to undertake this endeavor alone, and that JSTOR certainly would not be building storage facilities for this purpose. It also became clear that even under a consortium arrangement without construction of additional storage facilities, the "cost" of each incremental copy would be extremely difficult to estimate. For this reason, we decided not to pursue a formal cost optimization model and instead developed a method to find the Pareto frontier that provides a single decision-maker or a collection of decision-makers the opportunity to make tradeoffs considering both tangible and intangible costs on one hand and survival probability on the other.

In the course of this journey, we learned that when accurate data and costs are extremely difficult to obtain—and therefore the decision-maker(s) will have to make best-guess estimates anyway—providing a framework for thinking about the problem, a systematic approach for performing what-if analysis, and decision support that offers a range of solutions may be much more useful than a more elegant but rigid approach that is sensitive to mis-estimates of data or cost coefficients. Certainly, these are lessons that we already understood intellectually, but the challenges of estimating essentially all of the parameters in our models made us viscerally more aware that analysis and decision support tools need to be designed not only with the decision-maker(s), but also with the "fuzziness" of the data, in mind.

Acknowledgement

We are grateful for a grant from Ithaka that allowed us to perform this study. We thank Roger Schonfeld of Ithaka for introducing us to this problem and for facilitating access to data and experts in the library community. We appreciate the considerable efforts of John Kiplinger of JSTOR in providing us defect data for journal print copies and other critical information. The work benefited from the advice of Clifford Lynch of the Coalition for Networked Information, Tom Teper of the University of Illinois Libraries, Barclay Ogden and Charles Eckman of the University of California--Berkeley Libraries, and Ivy Anderson and Emily Stambaugh of the California Digital Library. We are grateful for comments from participants in the Chief Collection Development Officers' meeting at the American Library Association conference in Anaheim, California, and seminar participants at Queen's University (Kingston, Ontario). We also appreciate the assistance of Elizabeth Durango-Cohen in the development of the spreadsheet models.

REFERENCES

Atkinson, R. (2001). "CRL Collections Assessment Task Force Report," Focus XXI, 2, December 2001-January 2002, and online at http://wwwcrl.uchicago.edu/info/focus/Focus.htm.

Baker, M., Shah, M., Rosenthal, D. S., Roussopoulos, M., Maniatis, P., Giuli, T., and Bungale, P. (2006). A Fresh Look at the Reliability of Long-term Digital Storage. In *Proceedings of the 1st ACM Sigops/Eurosys European Conference on Computer Systems 2006* (Leuven, Belgium, April 18 - 21, 2006). EuroSys '06. ACM, New York, NY, 221-234.

Chopra, S. and M. S. Sodhi (2004), "Managing Risk to Avoid Supply Chain Breakdown," *Sloan Management Review* **46** (1), 53-61.

Chrzastowski, T. E., Naun, C. C., Norman, M. and Schmidt, K. (2007). Feast and Famine: A Statewide Science Serial Collection Assessment in Illinois. *College and Research Libraries* **68** (6), 517-532.

Hightower, C. and Soete, G. (1995). The Consortium as Learning Organization: Twelve Steps to Success in Collaborative Collections Projects. *The Journal of Academic Librarianship* **21** (2), 87-91.

Holley, R. P. (2003). Cooperative Collection Development. Encyclopedia of Library and Information Science, Second Edition, Marcel Dekker, Inc., 698 – 708.

Kleindorfer, P. R. and G. H. Saad. (2005). Managing Disruption Risks in Supply Chains. *Production and Operations Management* **14** (1), 53–68.

O'Neill, E. T. and Boomgarden, W. L. (1995). "Book Deterioration and Loss: Magnitude and Characteristics in Ohio Libraries," *Library Resources & Technical Services* **39** (4), 394-409.

Payne, L. (2005). Depositories and Repositories: Changing Models of Library Storage in the USA. *Library Management* **26** (1/2), 10-17.

Reilly, B. B. Jr. (2002). New Prospects for the Cooperative Preservation of Print Materials. *Resource Sharing & Information Networks* **16** (2), 151-158.

Roberts, E. P. (1988). Cooperative Collection Development of Science Serials. *The Serials Librarian* **14** (1/2), 19-31.

Rosenthal, D. S. H. (2008), "Bit Preservation: A Problem Solved?" Proceedings of the iPRES2008 Conference, British Library, August 2008.

Ruiz-Torres, A.J. and F. Mahmoodi (2007), "The Optimal Number of Suppliers Considering the Cost of Individual Supplier Failures," *Omega* **35** (1), 104-15.

Sarper, H. and Chacon, P.R. (2006), "The Reliability of Correlated Two-Unit Systems," *Annual Reliability and Maintainability Symposium—2006 Proceedings*, IEEE, Piscataway, NJ, 422-427.

Seaman, S. (2005). Collaborative Collection Management in a High-density Storage Facility, *College & Research Libraries* **66** (1), 20-27.

Seiden, P., Pumroy, E., Medeiros, N., Morrison, A., and Luther, J. (2002). Should Three College Collections Add Up to One Research Collection? A Study of Collaborative Collection Development at Three Undergraduate Colleges. *Resource Sharing & Information Networks* **16** (2), 189-204.

Sivathanu, G., Wright, C. P., and Zadok, E. (2005), "Ensuring Data Integrity in Storage: Techniques and Applications." StorageSS'05, 26-36, November 11, 2005, Fairfax, Virginia, USA.

Snyder, L. V. and Shen, Z.-J. M. (2006). "Supply Chain Management Under the Threat of Disruptions." *The Bridge* (National Academy of Engineering) **36** (4), 39-45.

Tang, C.S. (2005), "Perspectives in Supply Chain Risk Management," Working Paper, Anderson Graduate School of Management, UCLA.

Weitzman, M.L. (1998), "The Noah's Ark Problem," *Econometrica* **66** (6), 1279-1298.

APPENDIX: CALCULATING SURVIVAL PROBABILITIES FOR THE HYBRID MODEL

We can calculate the probability of survival for a given hybrid system (T, α, $n_{locked}(0)$, p_{locked}, $n_{backup}(0)$, p_{backup}) as follows. For each time period t, $t = 1,\ldots,T$, in succession, we identify the set of feasible states of the system, where the state of the system at time t is defined by the pair ($n_{locked}(t)$, $n_{backup}(t)$). For each feasible state in the given period, we can determine the probability of transitioning to each of the feasible states in the next period, which depends upon the total number of copies lost in the current period. (Recall that if the total number of copies remaining is at least $n_{locked}(0)$, we will retain $n_{locked}(0)$ in lock-up and any remaining copies are backups. If there are fewer than $n_{locked}(0)$ copies remaining, all of them will be in lock-up.) Thus, the state of the system can be completely determined from the total number of copies remaining

Each possible number of copies lost corresponds to a possible transition. Suppose the hybrid system is in state ($n_{locked}(t)$, $n_{backup}(t)$) and therefore the total number of copies in the system is $n_{locked}(t) + n_{backup}(t)$. After one period, the system can lose anywhere from $l = 0, 1, \ldots, n_{locked}(t) + n_{backup}(t)$ copies and each possibility corresponds to a transition to a new state of the system.

The probability of l losses in one period starting from state (n_{locked}, n_{backup}) (here, we drop the time subscripts for simplicity) is:

$$\sum_{x=0}^{n_{locked}} C_x^{n_{locked}} p_{locked}{}^x (1-p_{locked})^{n_{locked}-x} C_{l-x}^{n_{backup}} p_{backup}{}^{l-x} (1-p_{backup})^{n_{backup}-l+x}$$

where C_z^y is the number of ways we can choose z out of y items and is defined only for $z \leq y$. Each term in the sum is the probability that x of the locked-up copies are lost and l-x of the backup copies are lost. To derive the probability that l are lost, we need to sum over the feasible values of x (which are constrained directly by n_{locked} and indirectly by n_{backup}). We also note that for states of the form (n_{locked}, 0), the calculations are simpler, but conceptually, the process remains the same.

Now that we have a formula for the probability of l losses in a single period, we calculate the probably of l losses for each state of the system, (n_{locked}, n_{backup}), and l = 1,…, $n_{locked} + n_{backup}$. With this information, we can compute the probability of being in each possible state in the next period. For example, transitions to state (2, 3) in some period t could occur due to zero losses starting in state (2,3), one loss starting in state (2,4), two losses starting in state (2,5), and so forth. So the probability of being in state (2,3) in period t is the sum of the following terms: P{being in state (2,3) in period t-1} P{zero losses starting from state (2,3)} + P{being in state (2,4) in period t-1}P{1 loss starting from state (2,4)} + **….** (For states of the form ($n_{locked}(t)$, 0) where $n_{locked}(t) < n_{locked}(0)$, the preceding states also include those of the form ($n_{locked}(0)$, n_{backup}) where the associated transitions involve the loss of more than n_{backup} copies).

We repeat this process for successively larger values of t until we reach $t = T$. Having reached $t = T$, we add up the probabilities associated with all states in which at least one copy remains.