Counting the Costs of Digital Preservation: Is Repository Storage Affordable?

Stephen Chapman

Abstract

The Harvard University Library and the Online Computer Library Center, Inc. (OCLC) each manage centralized repositories optimized for long-term storage of library collections. Both organizations fully recover operational expenses by charging owners annual rates for managed storage services, regardless of materials use. The Harvard Depository assesses rates for analog storage per billable square foot. The OCLC Digital Archive assesses rates per gigabyte for storage of digital objects. Formats are significant, but not sole factors in determining preservation costs in these models. Owners' definitions of content integrity and tolerance for risk, which can change over time, are also important variables in the complex equation of preservation costs and affordability.

1 Introduction

Several years ago, Kevin Ashley of the University of London Computer Centre astutely observed that digital preservation costs correlate to the range of preservation services that are offered, not to the quantities or qualities of the objects being preserved. (Ashley, 1999) Preservation costs, he implied, would not necessarily be the same for identical collections deposited to different repositories under different agreements. To underscore these points, a taxonomy of digital preservation services—bitstore, minimal, full and optimal—has emerged in the literature, and first-generation repositories are carefully defining terms and conditions associated with various levels of preservation service. (LeFurgy, 2002) (NINCH, 2002) (HUL OIS, 2001) (DSpace, 2002)

Most cultural heritage institutions are likely to be consumers of centralized preservation services, rather than architects and managers of digital repositories. As consumers, librarians and archivists should carefully assess the obligations associated with any digital preservation pricing model, particularly the amounts of metadata and fees expected from them in order to ensure longevity for their collections. Repositories that manage ingest (acquisition), archiving and delivery services, for example, will have higher operational costs than ones solely dedicated to object management.

This article examines pricing associated with *one* component of digital preservation, repository storage, at one organization (OCLC), at one point in time (2003). OCLC's annual prices are compared to baselines (annual prices) to store comparable collections in analog formats in the Harvard Depository. Although the data presented herein are narrowly constrained, the model in which repositories bill content owners annually for preservation services should be familiar to librarians and archivists at many institutions. Organizations that pay for document management, underground storage or cold-storage services for archives, manuscripts, books, film and other media recognize that under the terms of these repository-owner agreements, longevity depends as much upon stability of funding (annual dollars) as stability of macro and micro environments.

Identifying challenges to finance repository storage sheds light on the significant challenge to make the broader spectrum of preservation services (from ingest to delivery) affordable. Repository storage is particularly important to quantify, then fund with reliable revenue streams, because it represents *ongoing* costs that apply to all materials designated to receive preservation services—however "preservation" is defined.

Several premises underpin the facts and figures about repository storage presented in the next sections:

- *The repository is the nucleus of preservation activity.*

  Four of the six "functional entities" in the *OAIS Reference Model*, including Archival Storage, pertain to the repository. Notwithstanding the LOCKSS model, to preserve through distributed replication, *OAIS* implies that without a repository, preservation is not possible. (LOCKSS, 2002), (OAIS, 2002)

  Used herein, the term "managed storage" encompasses at least three OAIS functional entities—Data Management, Archival Storage, and Administration—and possibly Preservation Planning as well. It is assumed that the content owners will delegate these responsibilities to a repository when possible, and that these functional entities are necessary foundations for all other preservation services.

  Emerging models for digital preservation reaffirm the fact that not all storage environments are equal. Optimizing environments, or sending materials to optimal locations, are proven affordable methods to minimize risk of obsolescence for large volumes of material. As James Reilly has succinctly observed, "Geography is preservation destiny." (Reilly, 2002)

- *The majority of content owners will be consumers of centralized repository services, not developers and managers of local repositories or digital asset management programs*.

- *Repository storage costs and payment schedules* (independent of costs for ingest or access) *must be affordable and manageable or content owners will withhold materials from deposit.*

- *Billing models and use patterns of existing (non-digital) centrally managed repositories are relevant indicators of what content owners (consumers) can afford to pay for managed storage services—independent of costs and benefits associated with retrieval.*

2 Managed storage costs in Harvard and OCLC repositories

The Harvard University Library and the Online Computer Library Center, Inc. (OCLC) each manage centralized, large-scale repositories dedicated to meeting the specialized storage and retrieval needs for cultural heritage institutions. The similarity of billing models in both repositories affords an opportunity to compare storage costs for collections in various formats, and to identify where cost gaps exist, why, and what remedies might be provided to make storage more affordable.

Established in 1986, the Harvard Depository (HD) is a climate-controlled media storage and retrieval facility that provides "…an extraordinarily high level of physical control, environmental protection, and inventory security [for] collections" in a cost-effective manner. (Harvard University Library, 2001)

Established in 2002, the OCLC Digital Archive is a storage and retrieval system and set of services that provide a standards-based, long-term solution for the secure storage of digital collections. (OCLC) As is the case for HD, OCLC services include secure data storage and management, ongoing administration and delivery of objects (by trusted applications), and object accounting and reporting services.

In exchange for the managed storage services provided by these repositories, object owners are billed at fully-loaded, annual cost-recovery rates. It is important to recognize that managed storage and "storage space"—i.e. square feet of constructed space, or GB of disk space—are not synonymous. HD and the OCLC Digital Archive are actively managed environments, where activities occur periodically at the macro- and micro-levels. Environmental control systems are regularly monitored and maintained at HD; staff also periodically inspect enclosures of some media, such as acetate film, to verify that higher risk items have not yet reached advanced states of decay. As explained below, OCLC activities occur most frequently at the file level, while macro-level activities such as media replacement ("refreshment"), software upgrades, and hardware replacement are scheduled at longer intervals.

For the purposes of this critique, prices and costs are considered to be the same. The HD and OCLC prices are structured to recover all expenditures associated with operating, maintaining, marketing, and growing large-scale repositories. Although expenditures may vary within each organization—OCLC, for example, serves a much larger market than Harvard—the billing models are comparable. (This is not the case for repository pricing in the Harvard University Library Digital Repository Service, where the $5.00 per GB recovers *only* the costs of disk space. Other key operational costs—such as data management, staffing, accounting and administration—are financed through Harvard's HOLLIS assessment and other sources.)

The Harvard Depository and the OCLC Digital Archive price their services as follows:

- *HD storage* is assessed at two rates per billable square foot (BSF). BSF represents a cubic dimension of 12" x 12" x 9". Current rates are $3.91 per BSF per year for standard climate-controlled storage (50° F, 35% Relative Humidity), and $9.85 per BSF for film vault climate-controlled storage (50° F, 25% RH).

- *OCLC storage* is assessed at three rates per gigabyte (GB), according to total amount of data deposited per account. Current rates are $60.00 per GB per year for 1-100 GB of data; $32.00/GB per year for 101-1,000 GB; and $15.00/GB per year for >1,000 GB. (Surface, 2002)

Notes

These billable rates for HD and OCLC storage *exclude* expenditures associated with preparing and depositing materials (ingest). HD charges additional fees to retrieve materials—on a per-item basis—whereas OCLC's per GB prices do include administrative access and delivery of deposited objects, but with the caveat that as technology changes, the objects that are retrieved may not continue to be "rendered" in human-readable form by web browsers and other contemporaneous software.

OCLC's current prices are for "bit preservation" services only. These include: data management and backup, ongoing virus and fixity checks, periodic media refreshment, disaster recovery, and support of administrative tools for owners to update metadata and generate reports. Prices have not yet been set for "full preservation," where OCLC would be obligated to provide standard bit preservation services, plus the capability to render intellectual content accurately, regardless of technology changes over time.) (Lavoie, 2000) (Surface, 2002)

OCLC rates per GB are calculated according to the cumulative size of the data objects being deposited, not the amount of storage media used. Due to storage redundancies essential for backup, each GB of deposited data requires more than 1 GB of disk space for repository storage.

Harvard content owners deposit large volumes of material to HD. By 1999, five additional storage modules had been added to keep pace with community needs. This rate of growth indicates that BSF rates have not yet presented barriers to deposit. Nor have increases to rates motivated owners to withdraw deposited materials. Excepting cases of temporary storage for building renovations, once collections are deposited to HD, they stay at HD. Thus, the BSF rate is a viable tool to budget long-term preservation costs for many paper and film media that have predicted life expectancies (based upon accelerated aging tests) in regulated storage environments.

Since 2002, OCLC has accepted deposits of data in specified formats (HTML, PDF, ASCII, BMP, GIF, JPEG, TIFF), with batch deposit services inaugurated in early 2003. Like the BSF rate at HD, the OCLC per GB rate is a useful budgeting tool. Acquisition specialists and other managers may calculate ongoing preservation costs for digitized and born-digital materials. By adding these to initial costs for purchase, processing (e.g., cataloging) and deposit, one may then estimate full life-cycle costs for stated retention periods. Curators interested in digitizing collections may weigh costs and benefits over the long-term by calculating the impact of various digitization specifications (particularly those related to file size) upon managed storage costs from digital repositories.

Cost studies positing that electronic storage is or soon will be less than traditional (brick-and-mortar) storage have used library construction and climate-controlled storage facility costs as baselines to measure the relative costs of digital storage. Lacking citable costs from operational digital repositories, authors of these studies have relied upon indicators of repository costs (e.g., media prices for storage) to advocate digitizing materials—particularly books—to minimize preservation costs. (Lesk, 1996) (Kingma, 2000)

How do the "real costs" for repository storage (that arise from Harvard's HD and OCLC's Digital Archive billing models) compare today? What do these comparisons reveal about the economics of digital preservation? Are we approaching affordability?

Tables 1 and 2 note the annual costs to store comparable collections in HD and OCLC.

Table 1. Annual Costs for Managed Storage at Harvard Depository, by Format

	Format	Quantity	# BSF	$ unit cost	Total
Text	books	2,202 vols	176	3.91/BSF	**$ 688**
	35mm microfilm	596 reels	43	9.85/BSF	**$ 424**
Photos	35mm color negative	3,000 images	0.5	"	**$ 5**
	4 x 5 color transparency	1,200 images	2	"	**$ 20**
Audio	¼" tape	40 hours	1	3.91/BSF	**$ 4**
Moving images	Motion picture film	20 hours	1	9.85/BSF	**$ 10**

Table 2. Annual Costs for Managed Storage at OCLC's Digital Archive, by Format

	Format	Quantity	# GB	$ lowest unit cost	Total
Text	ASCII with encoding	728,862 files	2.09	$15/GB*	**$31**
	1-bit page images	"	70	$15/GB*	**$1,050**
	8-bit page images	"	3,161	$15/GB	**$47,415**
	24-bit page images	"	9,484	$15/GB	**$142,260**
Photos	24-bit PhotoCD (~10.7 MB)	3,000 images	31.4	$15/GB*	**$ 471**
	24-bit TIFF (~200 MB)	1,200 images	268	$15/GB*	**$ 4,020**
Audio	96kHz/24 bit AIFF	40 hours	82	TBD	**--**
Moving images	"lossless" (62 Mbps)	20 hours	4,359	TBD	**--**

* these $15/GB prices assume that owners have already reached the 1,0001 GB threshold for their account; if these were first-time deposits, prices would increase to $32/GB for deposits of 101-1,000 GB and $60/GB for deposits totaling less than 100 GB.

As explained in the case studies below, the numbers for textual and photographic materials correspond to Harvard collections that have been reformatted in recent years, whereas the audio and moving images examples point to arbitrarily selected quantities of material—at file sizes consistent with high-quality ("lossless") reproductions—to illustrate the significant amounts of space these formats would occupy if deposited to digital repositories.

The total costs in these tables make two salient points:

- A *combination* of factors influence storage cost in traditional and digital repositories. These include: the repository's unit rate for billing, the type and number of media being deposited (printed books versus microfilm; printed books *and* microfilm; 1-bit versus 24-bit digital images), the number of versions being deposited, and the relationship between information content and media format (35mm versus 4 x 5 film; ASCII versus 1-bit versus 24-bit digital images). As Kevin Ashley has stated, "the primary influences on archival cost for a digital archive … are analogous with those which influence a more traditional archive or library." (Ashley, 1999)

- On the other hand, there are cases—such as audio and moving images—in which the pricing model to charge by size makes format *the* key cost variable.

Repository developers and administrators are keenly aware of two challenges: to make all storage more affordable, and to close significant cost gaps among various content and media types. Research and development efforts in these arenas are essential to model, scale and sustain digital preservation services—particularly in anticipation of scenarios in which owners are no longer able to pay for managed storage services due to changes in priorities, loss of revenue, or even the demise of the organization itself. Minimizing repository storage costs—for both analog and digital materials—increases the likelihood that others will "rescue" or continue to sustain content in jeopardy of financial as well as technological obsolescence.

3 Cost gaps

Unit prices for storage are powerful tools for preservation planning. By projecting storage costs for various formats, and considering associated benefits and tradeoffs, curators are empowered to develop collections-appropriate preservation strategies. Although it may seem contradictory, fixed rates (per BSF and per GB) do not necessarily dictate fixed costs. This costing model is flexible in practice, since curatorial decisions about content integrity and risk management ultimately determine what gets deposited, and therefore, what storage costs will be.

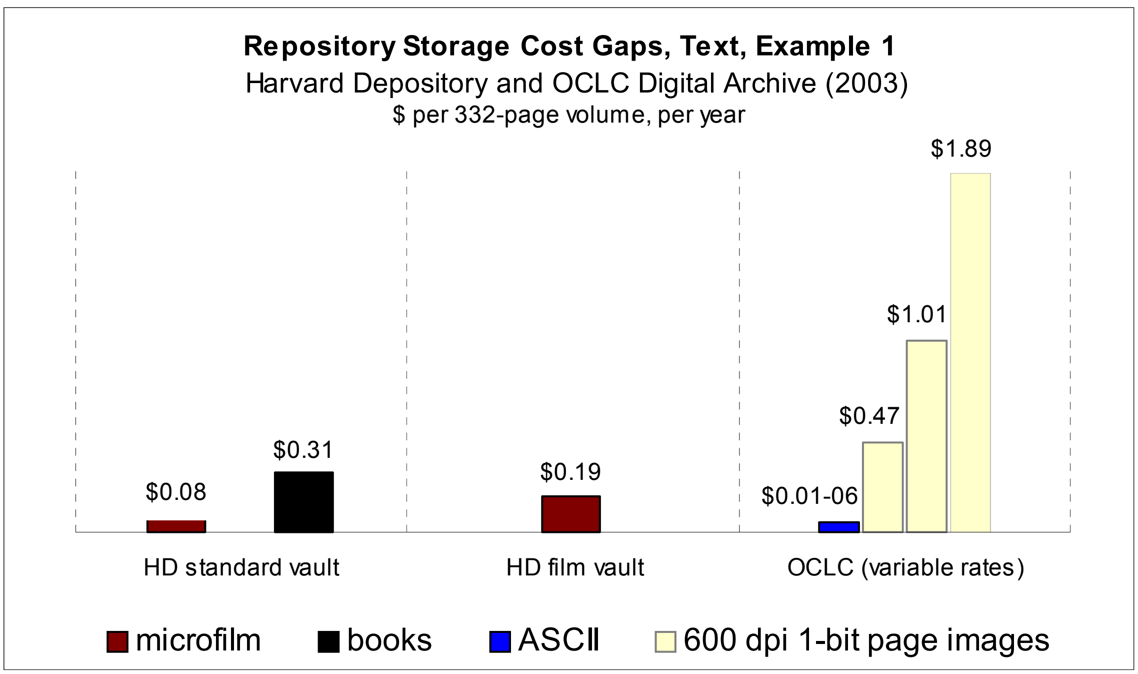
3.1 Case Study: Text Formats

With support from the National Endowment for the Humanities in 1998-99, Harvard microfilmed 2,202 volumes from the American Historical Textbooks collection at the Monroe C. Gutman Library, Harvard Graduate School of Education. Published between 1800 and 1940, these volumes contain approximately 729,000 pages. (Weissman, 2002)

Following microfilming, the Gutman Library retained the books in their library and deposited two generations of preservation microfilm (camera and print masters) to HD. The microfilm reels occupy 43 BSF in HD's film vault; if the original books (332 pages per volume) were also deposited, they would occupy approximately 176 BSF in HD's standard vaults. Neither the microfilm nor the books have been digitized, but estimates for digital file sizes of books with page dimensions of 5.75" x 9" are easily calculated per page, volume and collection. (See Tables 1 and 2 for calculated costs to store this entire collection, in various formats, in HD and OCLC.)

Chart 1 portrays the per-volume costs for managed storage of "text" according to several factors: format (microfilm, print, ASCII, or 600 dpi 1-bit page images with Group 4 compression), the pricing model of the repository offering managed storage service, and the choice of service within a repository—e.g., HD film vault versus standard vault storage. Note that these calculated annual costs per volume refer only to "masters" in cases such as microfilm or page images where delivery versions might be managed locally for more convenient access.

Chart 1. Relative Costs to Store Text "Masters":
Microfilm, Print, ASCII, and 600 dpi 1-bit Group 4 Page Image Formats



The microfilm storage costs in Chart 1 reveal that choice of HD storage (standard or film vault) is significant. The film vault's benefit of 10% lower Relative Humidity than standard HD storage comes at the expense of an additional $0.11 per volume per year, more than doubling repository storage costs for text stored as 35mm microfilm.

Decisions regarding the number of versions to deposit are important for risk management and preservation budgeting. Standard preservation microfilm practices, for example, yield two copies of film per volume: the camera negative and the second-generation print master. Although institutions could theoretically cut the microfilm storage costs presented in Chart 1 in half by depositing only the camera negative to a repository, no one does this in practice. These modest annual cost savings (maximum of $0.095 per volume) would be offset by the expense of increased risk. Eliminating a second copy, or keeping it out of a repository, enhances potential for information loss through theft, damage, careless handling or accelerated deterioration.

Organizations inclined to replace printed books with more affordable (if not necessarily more functional) reproductions would likely find microfilm, 1-bit page images, and ASCII to be relatively affordable when compared to annual repository costs for books. The decision to discard books, or more simply, not to deposit them to a climate-controlled repository, "frees up" $0.31 per volume to invest in the preservation of the microfilm or digital reproduction masters. As illustrated in Chart 1, at $0.31 per volume, an owner could purchase 1.63 years of storage for microfilm in HD, 5.17 years of storage for ASCII ($60/GB base rate), or 0.66 years of storage for 1-bit images in the OCLC Digital Archive ($15/GB discount rate). Conversely, deciding to deposit books to HD in addition to depositing reproduction masters in OCLC or HD adds $0.31 per volume per year to a preservation budget.

In 1996, Michael Lesk predicted that "…the costs of the digital and traditional library operations [would] cross over in about five years," and that electronic storage would offer a "major cost advantage" within ten. (Lesk, 1996) Seven years later, we observe in Chart 1 that this prediction appears to be true for ASCII, but that prices for managed storage of compressed 600 dpi 1-bit page images (100 KB/page) have not yet approached the "cross over" point to traditional (HD) storage. The most favorable digital-to-analog cost gaps for this digital format are 1.52:1 for page images versus books, and 2.47:1 for page images versus microfilm stored in the HD film vault.

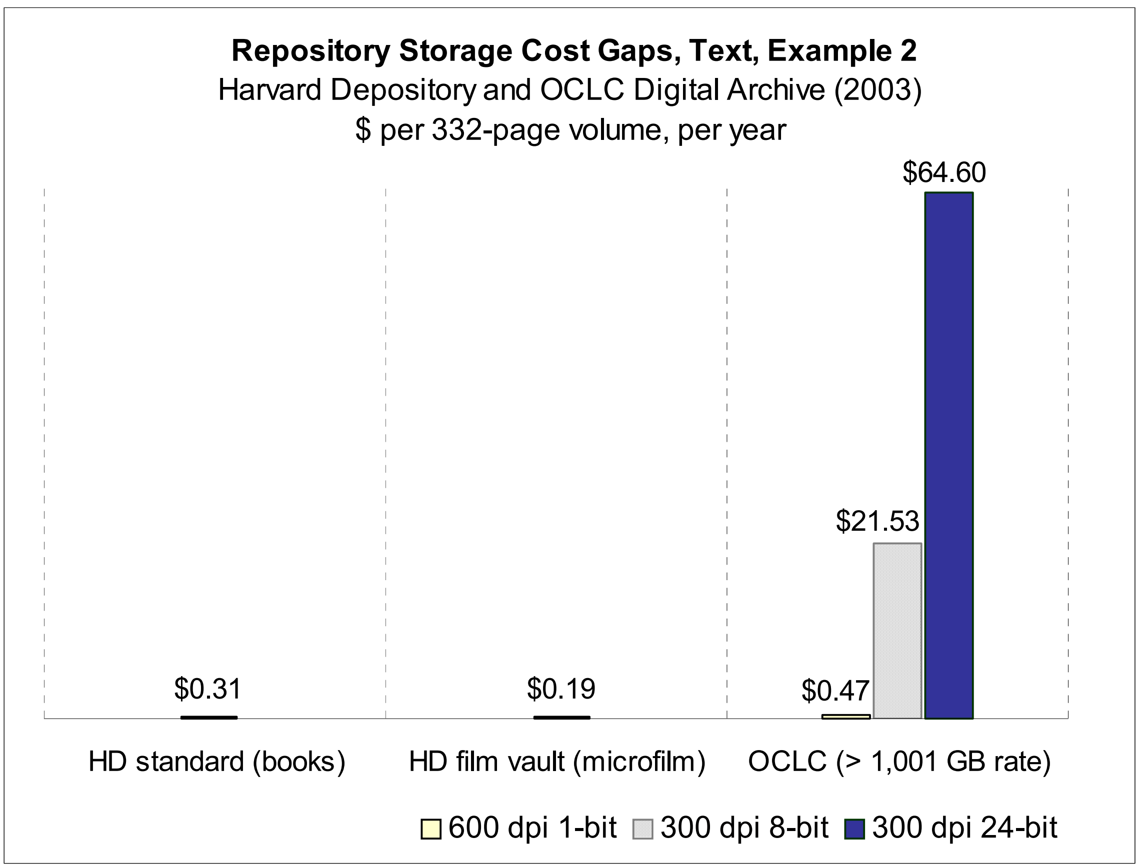
In the case of ASCII, by far the least expensive text format to manage in a repository, OCLC costs range from $0.014-$0.06 per volume, according to total volume of data for a given account. Note that volume discounts for ASCII ($0.06 per volume "baseline") would be reached at thresholds of 35.35 million pages (101 GB) and again at 349.93 million pages (1,0001 GB, 106,475 volumes). Encoded text formats (ASCII, XML, HTML) are extremely compact (3 KB/page) and have great advantages for searching, mark-up and display, but would not suffice as sole digital formats to reproduce historic illustrated textbooks.

The per-volume costs for 1-bit page images cited in Chart 1 all depend upon the use of lossless compression. Uncompressed, these file sizes would be approximately 22X larger, increasing the 1-bit to microfilm cost gaps to over 54:1, even at the OCLC $15/GB rate ($10.34 per volume for uncompressed 600 dpi 1-bit images versus $0.19 for microfilm).

Curatorial assessment of information integrity must be considered when calculating (or interpreting) repository storage costs for a given collection. Whether they are made at the point of digital acquisition or considered in reformatting workflows, these assessments are key to answering questions such as: "How many versions can an institution afford *not* to collect or preserve?” and “Are objects deposited as acquired or transformed (i.e., normalized) as a means to reduce versions and/or file sizes?” Curatorial assessments serve to define whether information content can be separated from its medium (format)—e.g., the text of the article is important, not the fact that it was originally formatted as a Microsoft Word file—or whether the medium *is* the content. Institutions are more likely to save multiple versions of content (textbooks and microfilm; Word documents and “normalized” RTF or ASCII), and therefore pay higher storage costs, when media are associated with meaning.

Finally, in addition to considering the relationship between content and medium, assessments of integrity must also account for quality. If digital texts were to be stored instead of printed books, for example, then *which* type(s) of digital would be viable—ASCII or page images, and if page images, 1-bit, grayscale or color? As Chart 2 illustrates, not all digital formats are equal (in size, information content, or storage cost).

Chart 2. Relative Costs to Store Text Masters
Microfilm, Print, 600 dpi 1-bit, 300 dpi 8-bit, 300 dpi 24-bit Page Image Formats



Charts 1 and 2 illustrate that costs are not necessarily fixed by pricing models, but are arrived at through a series of decisions having to do with quality, format (functionality), integrity and risk. They also underscore the importance of using correct terminology for the various formats that may be used to preserve the informational content of published works (such as textbooks). The imprecise term “digital books” makes each of the following statements ambiguous:

- One year of storage for a printed volume in HD ($0.31) purchases 5.17 years of storage for a digital book in the OCLC Digital Archive ($0.06). Printed books are 5.17 times more expensive to store annually than digital books. (Chart 1, books vs. ASCII)

- One year of storage for a printed volume in HD ($0.31) purchases 1.75 *days* of storage for a digital book in the OCLC Digital Archive ($64.60). Digital books are 208 times more expensive to store than printed books. (Chart 2, printed books vs. uncompressed 300 dpi 24-bit images)

It is essential to distinguish between text (e.g., ASCII) and page image formats (e.g., TIFF) when budgeting for digital storage. One also needs to beware of generalizing costs for one class of digital formats (e.g., 24-bit TIFF page images) without fully accounting for the possible range of file sizes within that format. For example, if the textbooks of our case study were scanned at 400 dpi (to match the University of Virginia’s practice in their Early American Fiction Project), file sizes and costs would increase by nearly 78% — for a total of $114.86 per 332-page volume per year.

Thus, while 35mm is a standardized size for microfilm and books can be grouped into similar size ranges for HD budgeting, beware of generalizing sizes for digital formats. For page images, variables among resolution, color content (black-and-white, grayscale, or color) and particularly compression all contribute significantly to file size, and therefore to billable annual costs in the OCLC model.

3.2 Case Studies: Non-Textual Formats

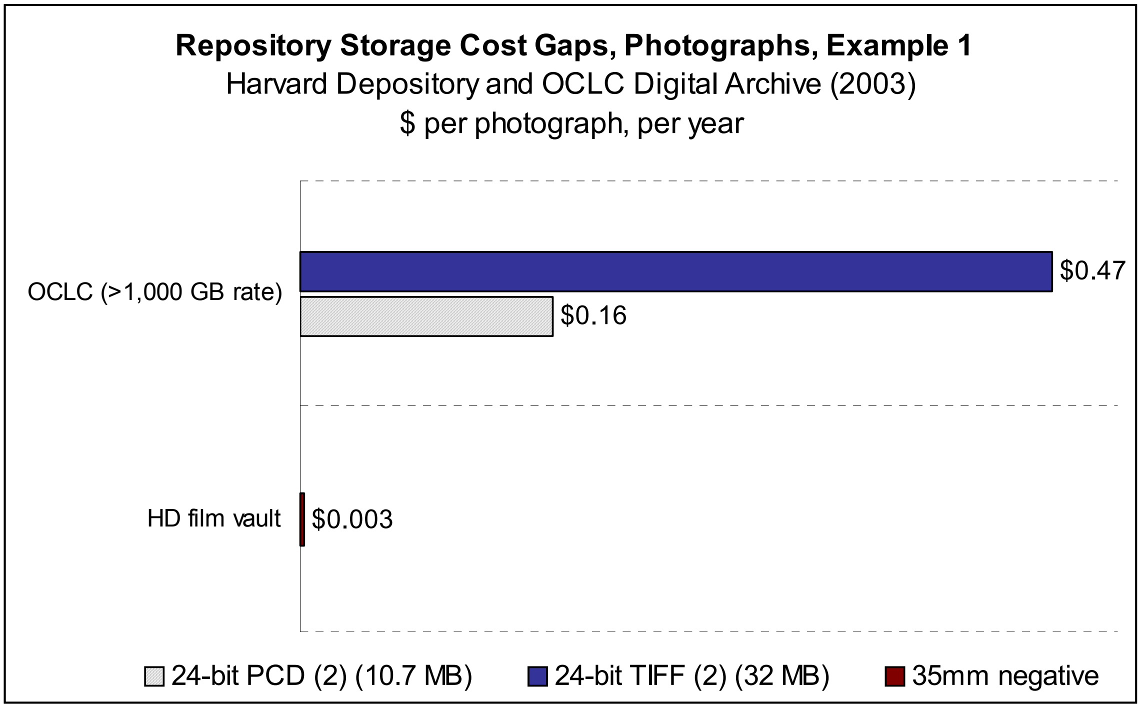
It is not surprising to find the cost gap between HD and OCLC storage to be significant for photographs, recorded sound and moving images. HD’s BSF rates are extremely favorable to the storage of these media, which, unlike books and printed archival materials, occupy relatively little space. Digital files of comparable quality or functionality to 35mm and 4 x 5 color film, 1/4" audiotape, and motion picture film occupy large amounts of space, particularly when saved in uncompressed formats.

3.2.1 Photographs

Storage costs for the two Harvard photograph collection case studies in Charts 3 and 4 highlight the economic tradeoffs associated with choice of format—within analog, and between analog and digital—in copy photography. With film-based photography, the per-image costs to store 35mm ($0.003) and 4 x 5 ($0.016) color negatives in HD’s film vault are both affordable to Harvard curators. Although they are not indifferent to 5:1 differences in storage costs for these media, it has been generally true that if an organization could afford to pay for photography, it could afford to pay for managed storage of the reproduction masters over the long term.

By comparison, the ratios of digital image-to-35mm film storage are 53:1 and 157:1 (Chart 3); and of digital image-to-4 x 5 film negative storage are 209:1 (Chart 4).

Chart 3. Relative Costs to Store Still Image Masters (Copy Photographs)
35mm negatives; Photo CD (PCD) and TIFF master images



The Printing and Graphics Arts Department, Houghton Library, Harvard College Library, recently photographed approximately 3,000 daguerreotypes to 35mm color negative film, then digitized the film to produce digital masters and delivery images. The calculated per-photograph cost to store the 35mm negatives in the HD film vault is approximately $0.003 per year (3,600 negatives per BSF). The digitization methodology in this project yielded two high-resolution masters: one that included grayscale and color bar targets in the frame beneath the photograph, and one "cropped master" optimized for batch production of delivery images. The OCLC prices in Chart 3 (at the $15/GB discount rate) account for the aggregate prices to store the two digital masters per original photograph: the shorter bar ($0.16 per photograph) represents the cost to store the 10.7 MB total of two Photo CD images; the longer bar ($0.47 per photograph) represents the 32 MB total to store two uncompressed 3,000 x 2,000 pixel TIFF images—had Houghton selected to use this format for the digital masters.

By selecting the Photo CD format for their digital masters, the Houghton Library would save approximately $0.31 per photograph per year if they deposited these two sets of masters to the OCLC Digital Archive. What costs are associated with this benefit of reduced storage price? The first is that image compression typically comes at the price of introducing irreversible quality losses to grayscale and color images, although these sometimes are not detectable to the eye until details are enlarged. The second is that someone—in this case Houghton Library under the implied terms of OCLC's "bit preservation" service—must monitor not only potential obsolescence of the format, but also a compression algorithm. In theory, migration or other transformations will need to be scheduled at shorter intervals than for uncompressed formats. MIT Libraries *DSpace* policies class formats as "supported," "known," and "unsupported," and Harvard's Digital Repository Service has comparable categories of Level 1, 2 and 3 formats. (DSpace, 2002), (HUL OIS, 2002)

Chart 4. Relative Costs to Store Still Image Masters (Copy Photographs)
4 x 5 negatives and TIFF master images

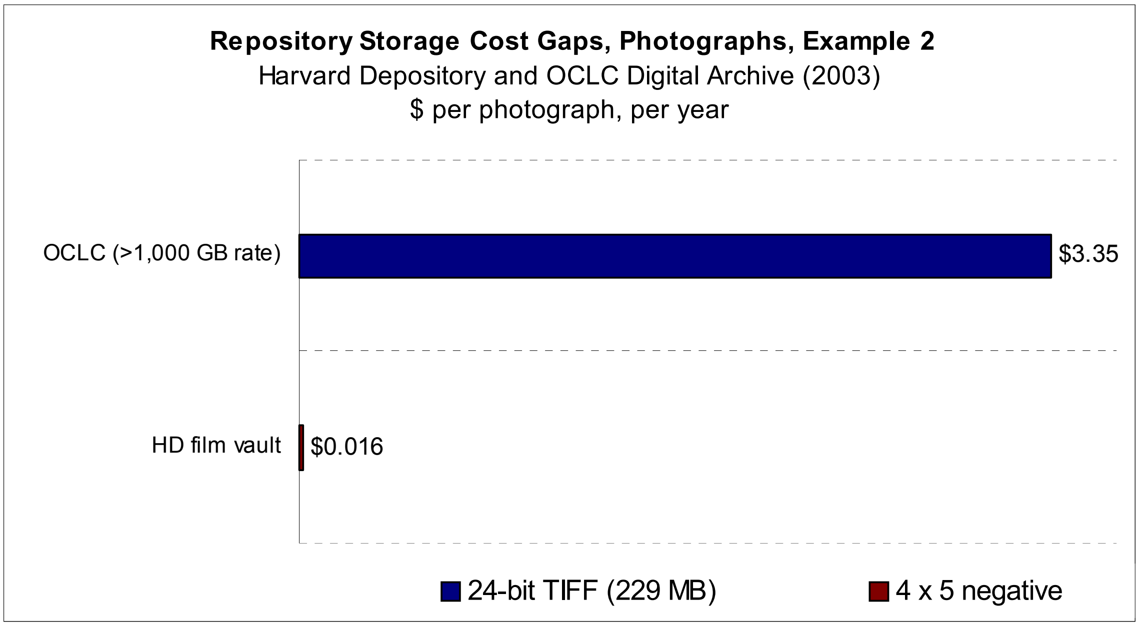


Chart 4 illustrates a second case study of a Harvard photograph collection. The Harvard University Art Museums photography studio closed its darkroom several years ago, making a complete transition from film to digital imaging for copy photography. In this project, they digitized approximately 1,200 vintage 4" x 5" negatives at 2,000 dpi 24-bit color. Average file sizes for the uncompressed TIFF images created at this specification are 229 MB. If deposited to the OCLC Digital Archive at the $15/GB discount rate, the per photograph costs are $3.35, versus the $0.016 price for managed storage of the "old model" 4" x 5" transparencies that would have been deposited to the HD film vault.

What remedies could be used to close these relatively large digital-to-analog gaps in repository storage for photographs? Three solutions that the museums considered, but rejected after evaluating image quality and risk management, would be to limit the size of files in digital photography (by scanning film at lower resolution); to use compression; or to keep camera masters outside of the repository, but deposit a reduced-resolution uncompressed version (downsampled from the master in an image processing program) as a preservation back up. (Note: the University Art Museums have deposited their 229 MB images to the Harvard University Library Digital Repository Service.)

A final strategy to adopt, which introduces the greatest preservation risk, is simply to withhold depositing any digital photograph to a managed repository until the assumed "inevitable cheaper storage technology" decreases unit prices, or until the providers of the repository service could be persuaded to price repository storage at units other than object size—such as number of items or rates of access (see, Ashley, 1999).

3.2.2 Audio and Moving Images

Tables 1 and 2 present file sizes of uncompressed high-density "preservation quality" digital audio (96 kHz/24 bit), and moving image formats of quality comparable to motion picture film. (Harvard's Eda Kuhn Loeb Music Library Audio Preservation Studio and other organizations are digitizing vintage sound recordings at these specifications today.) At OCLC's current per GB rates for other formats, the unit prices to preserve digital audio or moving images would be in tens ($30.76 for audio) or thousands of dollars per hour ($3,270 for moving images). (OCLC has not yet established prices for these formats.) Thus, the financial and technical challenges associated with long-term preservation of these formats are significant.

4 Conclusion

How do real costs for repository storage compare today? What do these comparisons reveal about the economics of digital preservation? Are we approaching affordability?

Managed storage costs represent only part of the full spectrum of digital preservation services. The case studies presented above document that there are many variables to the equation of managed storage costs. The nature of repository service is one key factor. At HD and OCLC, these prices correspond to carefully scoped obligations and services that ensure ongoing management of environments (HD) and data objects (OCLC), but *exclude* services to conserve or reformat items (HD) or to guarantee perpetual usability ("rendering") of digital objects delivered from the Digital Archive. The unit of billing for repository services is also important. Where organizations such as HD and OCLC bill annually for services based upon the size of material deposited/stored, object and file sizes become relevant, as do decisions about number of formats and versions to maintain in perpetuity. Under these pricing models, owners' decisions regarding acceptable levels of risk (regarding what they can afford to lose) and essential components of object integrity are instrumental to cost analysis.

Thus, managed storage costs are not fixed, but arrived at collection-by-collection by judicious decision-making. The choice of repository, the scope of service, the repository pricing model, and owner's decisions regarding formats, number of items, number of versions, and number of collections to deposit: all are potential variables, and therefore instruments, to negotiate for affordable prices for managed storage services from centralized repositories. These variables apply equally to traditional and digital repositories, and in both cases one potentially finds that some formats (content types) are more favored than others. A broad consideration of all of these issues—and topics for future papers—requires not only an assessment of cost variables, but also an accounting of benefits associated with these decisions. Managed storage makes many other services possible; perhaps peripheral services such as distributed delivery can be used to support some costs for repository storage. However, solutions must also be developed to maintain large quantities of material that are unlikely to be used or are used infrequently. We should consider storage challenges seriously and not assume that technology—such as "cheaper storage"—will provide easy solutions.

Acknowledgements

The author would like to thank Dale Flecker, Taylor Surface and Simon Tanner for their careful reading and commentary to earlier drafts of this paper.

References

Ashley, Kevin (1999) "Digital Archive Costs: Facts and Fallacies." In *Proceedings of the DLM-Forum on Electronic Records (DLM* '99), p. 123
http://europa.eu.int/ISPO/dlm/dlm99/dlm_proceed99_03.pdf (accessed 02.12.2002)

DSpace (2002) "Format Support," MIT Libraries & Hewlett-Packard Company, 2002
http://dspace.org/mit/policies/format.html (accessed 02.12.2002)

Harvard University Library (2001) "Harvard Depository: About the Facility"
http://hul.harvard.edu/hd/about-hd.html (accessed 02.12.2002)

HUL OIS Harvard University Library Office for Information Systems (2001) "DRS Policy Guide" http://hul.harvard.edu/ois/systems/drs/policyguide.html (accessed 02.12.2002)

Kingma, Bruce (2000) "The Costs of Print, Fiche, and Digital Access: The Early Canadiana Online Project." *D-Lib Magazine*, Volume 6, No. 2, February
http://www.dlib.org/dlib/february00/kingma/02kingma.html (accessed 02.12.2002)

Lavoie, Brian (2000) "Benchmarking the Preservation Process for Electronic Scholarly Journals," presentation delivered at DLF Fall Forum, November 19, 2000
http://www.oclc.org/research/staff/lavoie/dlf_1100.ppt (accessed 03.04.2003)

LeFurgy, William (2002) "Levels of Service for Digital Repositories." *D-Lib Magazine*, Volume 8, No. 5, May http://www.dlib.org/dlib/may02/lefurgy/05lefurgy.html (accessed 02.12.2002)

Lesk, Michael (1996) "Substituting Images for Books: The Economics for Libraries." presentation on 16 April 1996, in Symposium on Document Analysis and Information Retrieval, April 15-17, 1996 at Las Vegas, Nevada
http://lesk.com/mlesk/unlv/unlv.html (accessed 02.12.2002)

LOCKSS (2002) LOCKSS home page http://www.lockss.org/ (accessed 02.12.2002)

NINCH (2002) Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, and the National Initiative for a Networked Cultural Heritage (NINCH) "Preservation." In *The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials* http://www.nyu.edu/its/humanities/ninchguide/XIV/ (accessed 0.2.12.2002)

OAIS (2002) Consultative Committee for Space Data Systems (CCSDS). *CCSDS 650.0-B-1. Reference Model for an Open Archival Information System (OAIS).* Blue Book. Issue 1. January 2002. This Recommendation has been adopted as ISO 14721:2002. http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html (accessed 02.12.2002)

OCLC (2003) OCLC Online Computer Library Center, Inc. "Frequently Asked Questions: OCLC Digital Archive," http://www.oclc.org/digitalpreservation/services/archiving/digital/faq.shtm (accessed 03.04.2003)

Reilly, James (2002) "Preservation Strategy and Tools." Lecture delivered at *Preserving Photographs in a Digital World* seminar, Rochester, NY, August 22, 2002.

Surface, Taylor (2002) "Supporting the Digital Collection Life Cycle: OCLC Digital & Preservation Resources," presentation delivered at NELINET Digital Reality III Conference, October 2002 http://www.nelinet.net/conf/digital/driii_02/surface.ppt (accessed 03.04.2003)

Weissman Preservation Center, Harvard University Library (2002) "NEH Microfilming Statistics" http://preserve.harvard.edu/news/microfilm/nehstats.html (accessed 02.12.2002)