

**The Digital Book Project of the Oxford University, Cambridge
University Presses and the University of Pennsylvania Libraries**

Final Report

The Digital Book Project of the Oxford University and Cambridge University Presses and the University of Pennsylvania Libraries Final Report

Table of Contents

Executive Summary	i
Narrative History of the Project	
Motivation	1
Planning and Development Phase	2
Public Launch	3
System Upgrade	4
Focus Groups	4
Tools and Methods	
Digital Book Formats and Functionality	4
Search and Discovery Tools	5
Bandwidth	6
Summary of Digital Book Production	6
Data Capture for Assessment	7
Analytic Review	
The Collection	7
Library Processing	9
Unit Costs	11
Description of Use	12
Usage-Impact of the Digital Library	15
Reader Response – Overview	17
What the Readers Said	18
<i>Use of Digital Tools</i>	18
<i>The Project Collection</i>	19
<i>The Future of the Digital Book</i>	21
Summary Comments on Costs and Benefits	22
Conclusion	23

Table of Contents

Appendices

Financial Report	Appendix I
Documents Pertaining the Public Launch of the Project	Appendix II
Illustrations of Digital Book/PDF Functionality	Appendix III
Illustrations of Digital Book Web Site Functions	Appendix IV
List of Project Titles by Publisher	Appendix V
Letters of Agreement Between the University of Pennsylvania Library and the Presses	Appendix VI

List of Tables and Figures

Table 1. Collection by Area and LC Geographic Classification	8
Table 2. Annual Growth of the Collection	9
Table 3. Titles by Location of Source	9
Table 4. Time from Reception to Availability for Digital Titles	10
Table 5. Use of the Digital Book Collection by Network Domain	14
Figure 1. Time to Process a Digital File, Comparative by Publisher	11
Figure 2. Gap Between Receipt of Oxford Print and Digital Editions	12
Figure 3. Reader Sessions by Month	13
Figure 4. Frequency of Digital Saves (Virtual Circulation)	15
Figure 5. Annual Circulation of Print Books, Impact of the Project	16
Figure 6. Digital Saves in Relation to Print Circulation	17

Executive Summary

Over the past five to ten years, book-length electronic texts have caught the interest of scholars and librarians. However, the digital book continues to make an uncertain claim on the market for academic monographs and questions persist about its value and usefulness to scholarship. In the interest of learning how researchers at a major university would incorporate this new tool into teaching and learning, the University of Pennsylvania Library and the Oxford University Press sought the support of the Andrew W. Mellon Foundation to create a collection of digital books and study its use by academics in the humanities. The Penn/Oxford collaboration was joined in time by the Cambridge University Press and the affiliates of several Philadelphia area colleges.

From 1999 to 2004, the partners built a corpus of 772 digital books in history and allied disciplines. The books were cataloged and made searchable, as a collection and as individual units, on the World Wide Web. The collection attained sufficient mass to attract a following after some eighteen to twenty-four months. This allowed Penn a three year time frame to analyze and compare data on the use of digital and hardcopy, and interview readers in order to learn how academics interact with long texts in an online setting and integrate them into their work.

The study found that scholars use digital books in a cursory fashion, to check facts and to screen and evaluate content before committing time and possibly money to access hardcopy. The ease of finding and saving digital books and their unlimited availability induced high rates of “virtual” circulation. Nearly 50% of reader sessions included the download of a full digital book. The study found twenty downloads on average for the collection’s 772 titles. Data gathered during the study suggest online availability lessened demand on the library’s print collection. Still, readers made a deliberate effort to borrow or purchase print after examining specific digital editions. Use of the digital books was evenly spread across the collection, although the more frequently used titles in print tended to be the more frequently used titles online.

Digital book collections developed with PDF technology can be assembled quickly and at modest cost. The per title digitization cost identified in this study was about \$14.00 below the cost of print acquisition, when a consortial model for packaging and licensing a collection is assumed. The favorable cost implications of the digital book recommend it as a means of providing inexpensive, multiple copies of high-use materials and effective online access to reference sources that have not yet migrated to electronic form.

Focus groups of faculty and students provided substantial data on the interaction between readers and digitized texts. The consensus across the focus groups was unanimous that online access was desirable for search and text manipulation, but extensive reading will not occur online. As a result, the digital book is regarded as a powerful supplement but not a replacement for hardcopy text. In addition, student and faculty readers unanimously expressed a desire for online access to primary rather than secondary sources. Image-based materials are in particularly high demand. So is the ability to search and access content across data repositories.

Readers afforded lesser priority to facsimile display—PDF’s particular strength—than to features that more resemble database functionality. In general there is acceptance of text presentation in image form (PDF), but the study found an even keener desire among users for texts in encoded format. Encoded text is seen to provide better search capability and resolution, easier navigation, and higher quality image rendering than PDF.

Readers and librarians alike are enthusiastic about digital books because they promise new work and cost efficiencies. They also provide the potential means for resolving tensions within the current paradigm of scholarly communication. Younger faculty in particular hope they might be used to expand publishing opportunities that are beginning to narrow within the humanities, and librarians see a consequent benefit in the broadening of content for their collections and for future use by future scholars.

Final Report

The Digital Book Project of the Oxford University and Cambridge University Presses and the University of Pennsylvania Libraries

Narrative History of Project

In the mid 1990s, the emergence of book-length, full-text resources began to arouse significant interest within academic and commercial spheres. Scholars and librarians saw new and powerful tools for scholarly communication, research, and learning in the emerging technology. Publishers saw the potential for new markets and products, and significant implications for their production modalities and costs. The electronic or digital book raised important questions for all. For example: would faculty and their students use the digital book? How would it alter research behavior and teaching methods? Would it replace or merely supplement the printed codex? Would it create new market demands? How would it alter the workflow of libraries and publishing houses? And how would it be packaged and sold?

Motivation

Motivated by these and other questions, leaders of the University of Pennsylvania Library and the Oxford University Press came together late in 1998. Their purpose was to discuss a collaboration that could inform digital book development from a technical and an academic standpoint. That initial meeting ended with a plan to build a collection of monographs in electronic form and to study its production and use. The parties foresaw a need to give the collection a focus that complemented strengths of the University and the Press. History was the choice, because it represented an area where Penn faculty and programs have international stature, and Oxford has a wide-ranging and established annual output.

Oxford projected a steady stream of content from its history list of 300-500 new titles per year. The Library had the computer capability and web design capacity to provide a reliable service infrastructure and production environment.

The Penn community offered a sizable test bed for the study of digital book use. However, early in the planning the Library and Press agreed to expand the project audience to the students and faculty of the Bryn Mawr, Haverford, and Swarthmore Colleges (the TRI-CO colleges), a group of excellent liberal arts schools with a common online library catalog and close institutional ties, among themselves and with Penn. Collectively the participating schools would provide an audience of approximately 225 faculty in history and allied disciplines, and as many as 1,000 students in the graduate and undergraduate ranks. While this group would form the collection's natural user base, the digital books would also be available to the wider faculties and student bodies of the four schools. In addition to enlarging the audience, the mix of schools would also offer the study contrasts in programs and institutional missions.

Planning and Development Phase

The Mellon Foundation agreed to fund the project in April 1999. By the following year Oxford had provided seventy-four books. The Library had installed a server, developed a workflow for book processing, hired graduate student interns to process the digital files received from the Press, and developed a web site to host the collection. Penn and Oxford had an interest in raising general public awareness of their efforts. For this purpose, Penn also created a public preview site that showcased the collection and digital book functionality. The preview site hosted three texts, which were free of copyright encumbrances, as demonstration samples.

In the planning phase of the project, the Library retained Dr. Malcolm Getz of Vanderbilt University to oversee project assessment and analysis.

Throughout the project, Penn received and cataloged a print copy of each digital book through the standard approval process used to acquire history titles. Each catalog record contained a link to the full text on the web, and these records were shared with the TRI-CO colleges unified catalog system.

At this stage, plans were in place to launch publicity and heavily advertise the collection. That effort had to be delayed until the number of digital books achieved a critical mass--a threshold the project did not cross until the late summer of 2001 when the collection approached 300 titles. In the meantime, the Library posted a link to the digital book collection on its homepage to begin raising the project's visibility. The digital book web site included a press release and a page describing the intent and goals of the Penn/OUP collaboration.

During the build-up period the Library also announced its partnership with Oxford in the library and general media, and among selected audiences at Penn and the "TRI-CO colleges", including faculty advisory groups, history departments, and graduate student organizations. One response to this consciousness-raising effort came from the Cambridge University Press (CUP). Like Oxford, Cambridge was interested in data that could inform its digital book publishing plans. By the start of the 2001 fiscal year, CUP and Penn were working on a letter of agreement that would add Cambridge content to the collection. That agreement was finalized in September 2001, and seventy-five CUP books entered the collection that November.

Public Launch

Fall 2001 also marked the formal roll out of the collection with much on-campus publicity and a series of discussion and demonstration programs for faculty and students. Over the following months the Library worked to keep the collection in the public eye. Features ran in the *Daily Pennsylvanian* (Penn's student newspaper), and the *Pennsylvania Gazette*, a popular campus magazine read by faculty, staff, and alumni. At faculty and curriculum meetings, and in one-on-one sessions with faculty, librarians promoted the collection within humanities departments. Late in 2001, librarians also began sending faculty and graduate students periodic email alerts containing hyperlinks to new titles. (Materials pertaining to the public launch appear in Appendix II)

System Upgrade

Early in the spring of 2002, Penn's Digital Projects Librarian completed a major upgrade of the project web site. New search functions were introduced, including subject and series title searching, and Boolean searching with parameters for limiting by time period and geographic location. New displays allowed users to hotlink on Library of Congress Subject Headings, and scan other bibliographic details derived from catalog records. The site redesign also incorporated a link to the Oxford and Cambridge web sites to purchase books; in Oxford's case the purchase option included a modest discount on the list price.

Focus Groups

In the eighteen months between the formal roll out in fall 2001 and the start of spring term 2003, the collection had grown from 300 to nearly 700 titles. Though well below the anticipated number of books, the collection covered a broad range of scholarly interest. A reasonably large number of books had been available over a long enough period for faculty and students to explore and begin integrating the digital book into their work. This assumption was supported by usage data, which are further described below. At this phase of the project, the Library assembled focus groups at both Penn and at the TRI-CO colleges to gain insight into faculty and student views of the service. A summary of the focus groups appears below.

Tools and Methods

Digital Book Formats and Functionality

Word processing, computer assisted publishing, and new off-set printing technologies have made modern book production a digital process. From the point of composition to design and page formatting to the creation of galleys and plates, a book passes through a series of digital manifestations before it rolls off the press.

At least until recently, the digital files that went into printing a book generally ended up in archives or may even have been discarded. But by the mid 1990s, the digital byproducts of book-making—postscript documents, desktop publishing files, and other

forms of electronic media—were being viewed with new interest thanks to technologies that could convert them into web-accessible content, technologies like the Adobe Portable Document Format (PDF). With tools such as PDF, publishers acquired a means of building complete digital books in a fraction of the time image or OCR scanning required, and without the cost, complexity, and inherent errors of scanning.

At an early stage of planning, Penn and Oxford seized on PDF as the tool that could enable them to build a digital library quickly and cost-efficiently using the unrealized capacities of the Press's postscript archive. The PDF format offered a range of potential benefits that the project would both leverage and analyze. On the production side, it required inexpensive, off-the-shelf applications, and a modicum of computing power. It produced a digital facsimile of the printed work, which retained the design and organization of a codex. PDF allowed for digital and print editions to have identical pagination, indexing, text and image adjacencies, and footnote presentation. While preserving the design values of print, PDF offered computer functionality, such as word search, copy-and-paste, and the ability to print and download. And finally, as an established medium for accessing text with a web browser, PDF had become a ubiquitous means of content delivery, one that lowered the technical hurdles readers might face.

Components of Adobe Acrobat were used to develop the monographs in the Penn/OUP/CUP collection. Enhancements were added to the books using another commercial application, Compose from Infodata. Compose features included an exploding table of contents with quick links to chapters and subchapters, and links between index entries and corresponding passages in the body of the book (see Appendix III for examples of digital book layouts and functions).

Search and Discovery Tools

Within each digital book, the PDF reader (Acrobat) supported word search. To provide cross-collection searching, the Library used Verity software. Verity indexed the entire PDF and associated MARC metadata to enable full text, author, title, series, ISBN, and publication year searches. Verity scripts dynamically built search result pages displaying

author, title, book jacket image, and related links. ISBN control numbers linked MARC records in the Library catalog to each PDF, thus creating two-way connections between the catalog and project web site.

In addition to federated collection searching, the digital book web site design at Penn offered a number of browse features. Readers could quickly page through lists by author, title, series, and publisher, or view a list of the latest books in the collection. The browsing displays included summary bibliographic information, a link to the catalog for print availability, and a link to the publishers' web sites for more information or purchase options. Sample pages are provided in Appendix IV.

Bandwidth

To optimize file transfer across the range of possible network connections, the Library employed byte streaming. Any one session with a book involved multiple requests against the site's web server as readers browsed chapters or searched terms within a "volume". Byte streaming also provided signatures within the web server logs to assist in counting the number of views in a typical session and determine when readers saved entire books to their personal computers.

Summary of Digital Book Production

The digital books reached the Library in batches of varying quantity and at irregular intervals. Often the lag between shipments was long. Most Oxford books arrived on CDs from points around the world, and were usually accompanied by galleys. Cambridge items were sent via FTP from the United Kingdom. Postscript was the most common file format received, but files also arrived in PDF and, at times, as application files such as Quark and PageMaker, two forms of desktop publishing software. Raw book files arrived in parts corresponding to chapters. A project intern distilled these segments into PDF using Adobe Acrobat editing software, and then assembled the segments into a single large PDF document for inclusion in the collection. Occasionally the process required image scanning to add illustrated matter omitted from the source files.

The final PDF was enhanced with internal bookmarks and linkable tables of contents and indices using Compose. A book jacket image often came with each file, and when it did not, the Library scanned or manually created one. The image was then linked to the web site as a graphical aide-memoir. This would complete the book creation process.

The intern would then mount the fully prepared book file on the project web server and notify a cataloger who would create a bibliographic record for the OPAC. Print and digital editions of the same titles received distinct MARC records, usually based on the print version. The records for digital books had additional fields to allow for easy extraction in the event of future processing and for periodic distribution to the TRI-CO colleges for inclusion in their library catalog.

Data Capture

To capture usage data, the Library established server logging routines. Each night logs from the machine hosting the project web site were extracted into a storage array to be processed on a monthly basis. The Library captured IP address information to differentiate on- from off-campus traffic, and sort out Penn use from use by affiliates of Bryn Mawr, Haverford, and Swarthmore. Monthly tallies of digital book use were posted title-by-title in the Library Data Farm. The reports included session and view counts and counts of monthly downloads, that is, retrievals of entire monographs to the reader's desktop. The Library also compiled historical circulation data for Oxford books dating from 1997. These statistics would provide a baseline for analyzing the impact of digital book availability on the demand for print editions.

Analytic Review

The Collection

The Digital Book Project comprised a small-scale collection of 772 books in history from Oxford and Cambridge. Within the discipline, the range of topics was broad, covering political and social history of Europe, history of religions, classics, economic history, women's studies, history and sociology of science, historical anthropology, and other

areas of cross-disciplinary interest. Every major time period was represented, with strength in twentieth century history, history of the ancient world, and the Middle Ages. In terms of geography, European history, particularly history of England, France, and Germany, and U.S. history dominated the collection. There were also notable concentrations in history of India and Russia (see Table 1). The full list of titles appears in Appendix V.

Table 1. Collection by Area and LC Geographic Classification

	Number of Titles	Pct	Cml Pct
Europe	405	52.5%	52.5%
North America	164	21.2%	73.7%
General	94	12.2%	85.9%
Asia	56	7.3%	93.1%
Africa	20	2.6%	95.7%
Australasia	12	1.6%	97.3%
South America	5	0.6%	97.9%
Commonwealth Countries	5	0.6%	98.6%
Intercontinental Areas (Eastern & Western Hem	9	1.2%	99.7%
All Other Areas	2	0.3%	100.0%
Total	772	100.0%	

Overall, the project fell well short of its collection goals. Oxford had promised to deliver digital versions of nearly all of its titles in history, worldwide, for a five-year period. The total anticipated was between 1,500 to 2,000 titles. The Press delivered 640 titles in all, and of that number 624 were suitable for processing and made available to the public. The Cambridge University Press, whose participation wasn't contemplated at the outset, delivered 150 titles; of that number 148 were made available to the collection and two were withheld for technical problems.

Table 2. Annual Growth of the Collection. (Project years ran from April to March)

	CUP Mounted	OUP Mounted	Cumulative Total Mounted	Held Back	Annual Total
Year 1		74	74	0	74
Year 2		172	246	3	175
Year 3	94	247	587	11	352
Year 4	52	86	725		142
Year 5	2	45	772	0	47
5-Year Total	148	624		18	790

Titles came from eight Oxford sites around the world as shown in Table 3. Some of the Press's difficulty in supplying titles and some of the technical difficulties in working with the raw files arose from coordination problems among these distributed production facilities.

Table 3. Titles by Location of Source (mount books only)

	CAM	OUP							Total	Cumulative Total
	AU	AUS	CAN	INDIA	NC	NY	UK			
Year 1					38	4	32	74	74	
Year 2	10				50	27	85	172	246	
Year 3	94		2	9	7	37	48	144	587	
Year 4	52					33	21	32	138	725
Year 5	2					2		43	47	772
5-Year Total	148	10	2	9	7	160	100	336	772	

Note: AU = Austria, AUS = Australia, CAN = Canada, INDIA = India, NC = North Carolina, NY = New York, and UK = United Kingdom.

Library Processing

The Library did succeed in converting the files delivered by the two university presses into the files necessary for the digital library in a cost-effective and timely manner.

Shipments were punctuated by lengthy delays, with titles arriving in batches that caused

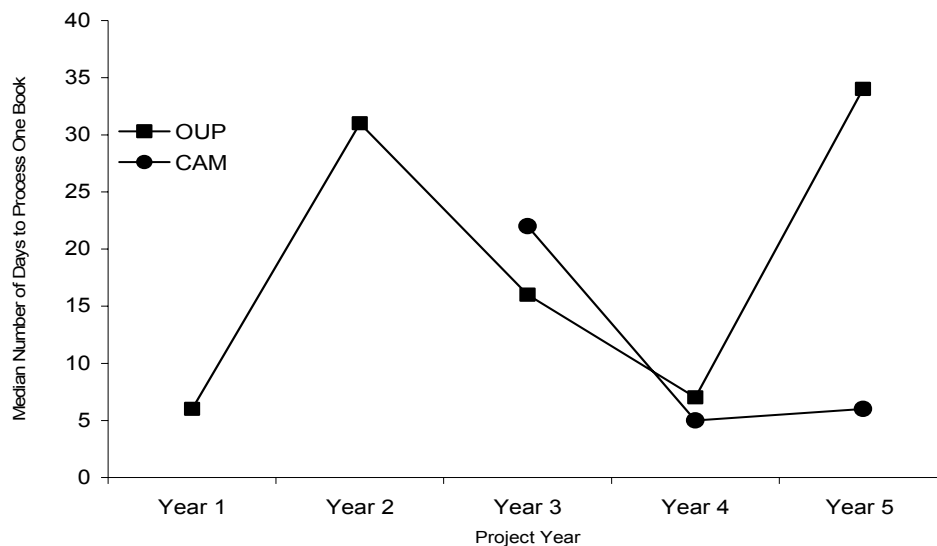
temporary processing backlogs. This led an average delay of sixteen days from the time of receipt until new titles became available to library users. Nearly 45 percent of the titles were available within two weeks of receipt and more than 70 percent were available within one month (see Table 4).

Table 4: Time from Receipt to Availability for Digital Titles

Processing Frame	CAM	OUP	Total	Pct	Cum Total	Cum %
1-5days	47	159	206	26.7%	206	26.7%
1-2wks	26	112	138	17.9%	344	44.6%
2-3wks	13	82	95	12.3%	439	56.9%
3-4wks	40	65	105	13.6%	544	70.5%
1-3mo	18	147	165	21.4%	709	91.8%
3-6mo		25	25	3.2%	734	95.1%
>6mo	2	17	19	2.5%	753	97.5%
>1yr	2	17	19	2.5%	772	100%
Grand Total	148	624	772	100%		

Digital files varied widely in quality and format, with problem files further extending output time. The most difficult problems involved labor-intensive manual fixes, such as image rescaling, page renumbering, and manual table of contents linking. Exceptionally long books had to be treated as two volumes, adding complexity and cost. Overall, eighteen of the 790 titles that were received as raw files imposed technical challenges the Library could not overcome. All but two of these were Oxford titles, and the majority of problematic files originated with Oxford as well. Cambridge consistently delivered titles in PDF format, and on the whole handled throughput more expeditiously than Oxford. In the end, more than 97% of the project books were mounted, although in some cases desired functionality, such as internal index links, had to be sacrificed. Figure 1 compares the rate of processing time by publisher over the term of the project.

Figure 1: Time to Process a Digital File, Comparative by Publisher

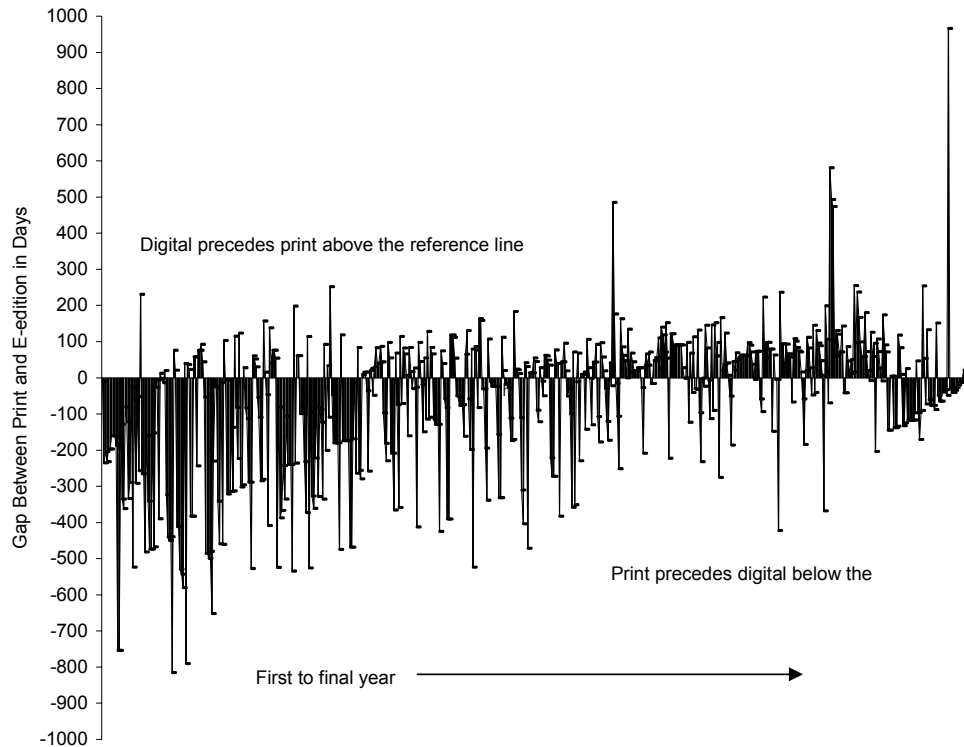


The Library received digital files from Oxford fifty-six days after receipt of the print title on average. (Cambridge was not considered in this assessment because it came into the project late and began with a large, retrospective consignment of books.) As Figure 2 indicates, printed items reached the Library ahead of their digital counterparts over the life of the project, although the gap tended to narrow as time passed. This change was presumably the result of Oxford’s effort to make workflow adjustments and improve the handling of digital assets generated in the print process.

Unit Costs

The Library incurred an incremental expense of \$94.24 per title to process raw digital files into the digital collection. A fixed cost in hardware, software, and system creation amounted to an additional \$150,000. Were the Library to process 1,000 titles per year for four years before replacing equipment and software at a leased value of \$150,000, the fixed cost would then be about \$50 per title. The cost of processing the digital files might be lower still if the raw files from the publishers were in a uniform format so that the work could become routine. And were the publishers to offer the digital collection to many libraries, the cost per title per library would be much less than the cost shown here. For example, if the variable cost of \$94 and an allocated fixed cost of \$50 per title created a digital collection shared by just

Figure 2: Gap Between Receipt of Oxford Print and Digital Editions



ten libraries, the processing cost per title per library would be about \$15. Penn’s processing cost for a print title averaged \$28.86 in fiscal 2003. Granted, print involves additional cost for adding storage capacity and additional handwork in circulation and shelving that is largely avoided with an all-digital approach. But on average, processing digital production files from publishers into full-text digital library documents appears to offer significantly lower costs for libraries than the processing and storage of print monographs.

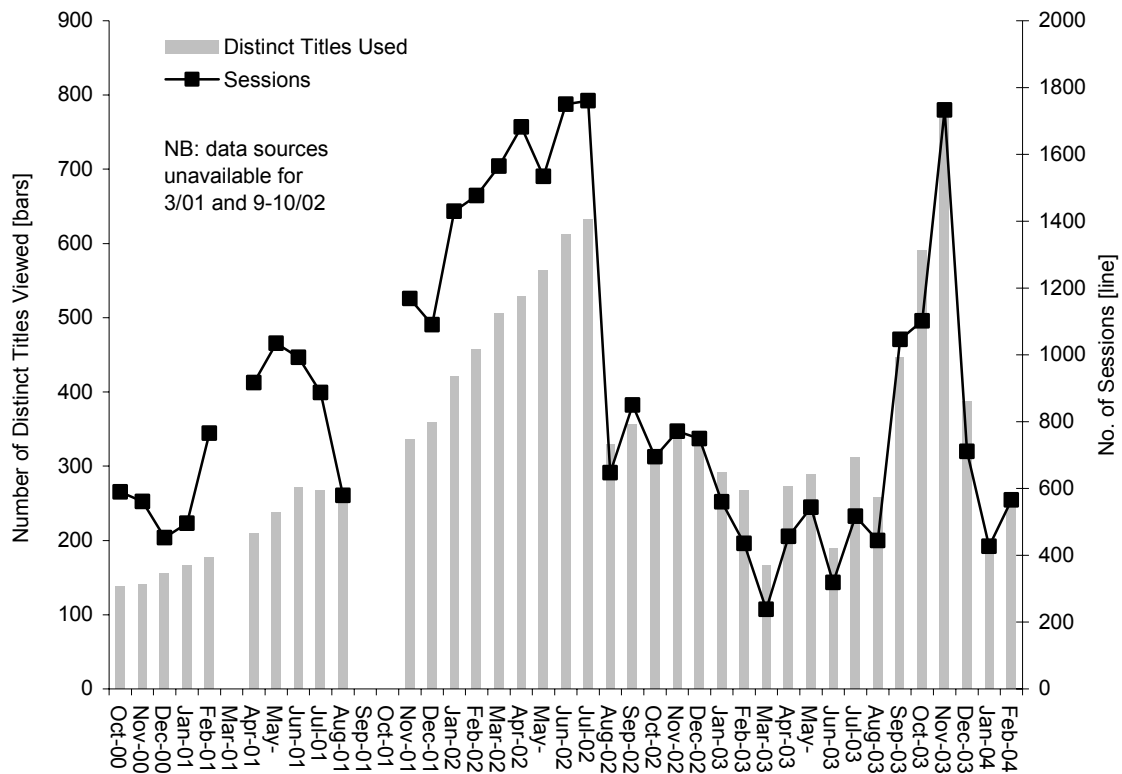
Description of Use

Data on collection use was gathered from Library server logs over a forty-one month period. In that time frame, the digital book collection hosted 31,749 reader sessions, where a session is defined as a successful connection with one or more books within a thirty minute time out frame. A thirty minute period of inactivity terminates the session count for a connection. Successful connections were marked in the Library’s web server

logs with Return Codes of 200 or 206. The usage statistics cited in this section exclude activity from staff computers, search engine indexers, and the project’s public access page.

Monthly session counts averaged 774 with every book in the collection used. Figure 3 shows the monthly distribution of reader sessions along with the number of titles used.

Figure 3. Reader Sessions By Month



Sessions could be initiated in a variety of ways, including clicking on a URL from the project web site, selecting a URL found in the Library’s online catalog, bookmarking a title within the web browser, or finding a link to a project text on a University of Pennsylvania web page or course site. The majority of sessions originated from the non-library, network sub-domains of the Penn campus. Table 5 provides a breakdown by principal locations of access.

Table 5. Use of the Digital Book Collection by Network Domain

	No. of Reader Sessions	Pct.	Cml. Pct.
Penn Schools and Computer Labs	19,280	61%	61%
Public ISPs used by affiliates of the project schools	6,543	21%	81%
UPENN Library	2,485	8%	89%
TRI-CO Colleges	2,168	7%	96%
Student Residences at Penn	1,273	4%	100%
Total	31,749	100%	

As defined here, a reader session is a close analog to perusing a book on the shelf, or browsing. The analog of borrowing a printed volume is the process of downloading or saving an entire PDF file to one's personal computer. Saves required a stronger commitment to using a text than browsing a few pages.

As stated earlier, the Library enabled byte streaming to optimize available bandwidth. The byte streaming process continuously transmits small portions of a large PDF to the web browser until the entire file is delivered or the reader terminates the session. Typically, the digital books transferred in packets of 3.2 kb. If the reader used the Acrobat "save" function to store a book on a local computer, the server bypassed byte streaming and downloaded the entire PDF in one transaction. This event appeared in the server log with a byte count equal to the file size of the downloaded book. These byte factors provided convenient signatures in the logs for measuring browses and saves. Over the 41 months of usage analysis, 46% of sessions resulted in a save; that is equivalent to 14,688 virtual circulations. The data derived from byte streaming show that the average reader session consisted of 9.5 file transfers after the initial connection. Based on the size of the individual digital books, the 9.5 average indicates extremely cursory use of the texts online. These data accord well with findings of focus group session presented elsewhere in this report. Specifically, readers employed the web to scan content, and then downloaded entire books for more detailed use, which usually included printing.

Of the 772 titles in the collection, all but one had at least one save, and the average rate of saves was nineteen over the forty-one months measured. This stands in marked contrast to the average rate of print circulations of the same titles, 0.33, over the same time frame. The ease of finding and saving files, and the unlimited availability of digital editions induced a higher rate of use than does print circulation.

Figure 4. Frequency of Digital Saves (Virtual Circulation)

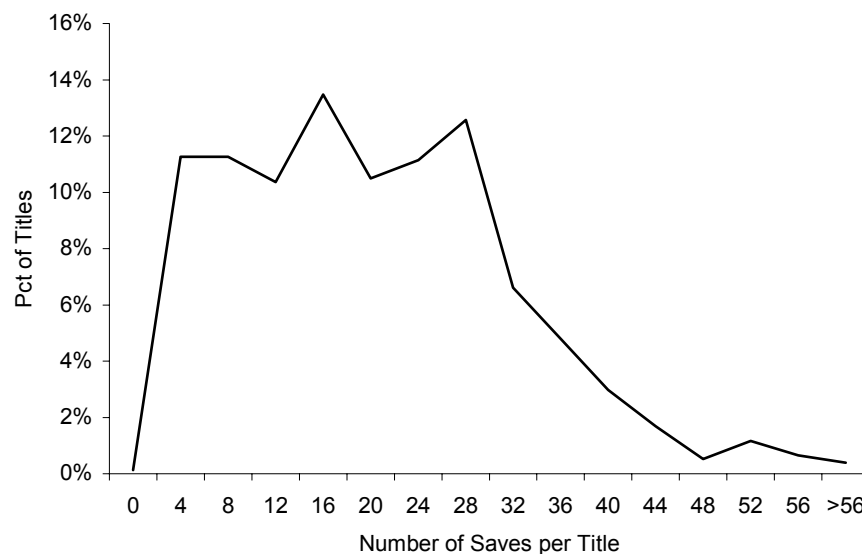
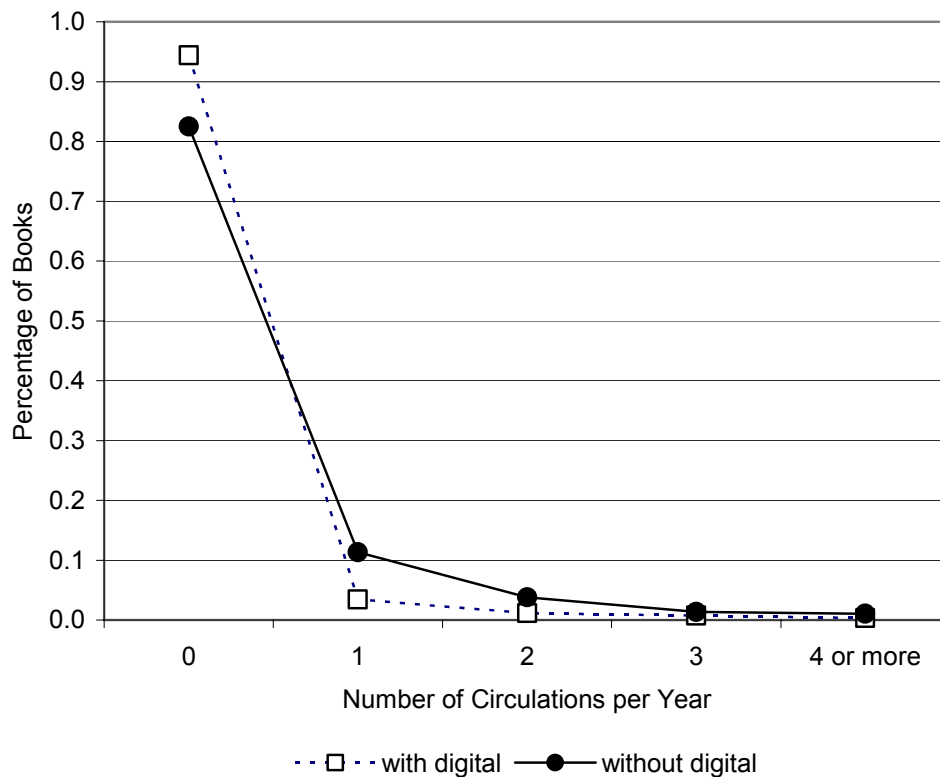


Figure 5 shows the distribution of saves for the 772 titles in the digital collection. A preponderance of titles had between eight and twenty-eight saves, and the range of saves for an individual title was as high as eighty-eight. From the point of view of utilization, the digital library appears to be a substantial success.

Usage-Impact of the Digital Library

The University of Pennsylvania Libraries tracked the circulation of titles in print from OUP and CUP. Figure 5 compares the annual rate of circulation of print books from OUP and CUP for titles that were available only in print with those that were available in both print and in the digital library. The figures for the print only circulation refer to the years 2000 to 2003 with a total of 40,711 titles in the end. The print circulation when digital is available refers to 2002 and 2003 with 735 cumulated titles. The Figure shows,

Figure 5: Annual Circulation of Print Books, Impact of the Project



and a chi-square test confirms, that the pattern of use of printed volumes differs with the introduction of the digital service.¹ The percentage of print titles that do not circulate in a year rose from 83% with print only to 94% with print and digital. With digital titles available, print use dropped. The average annual circulation for print only titles from OUP and CUP was 0.29 while the rate of circulation for print titles that were also available digitally was 0.10.² The print with digital titles are in history and the full set includes many disciplines, a possible source of difference here. However, the effect here may well understate a longer run effect. With a larger collection of digital titles, readers would more frequently consult the database for digital materials. The digital library contained primarily new titles while the print-only collection contained a larger share of

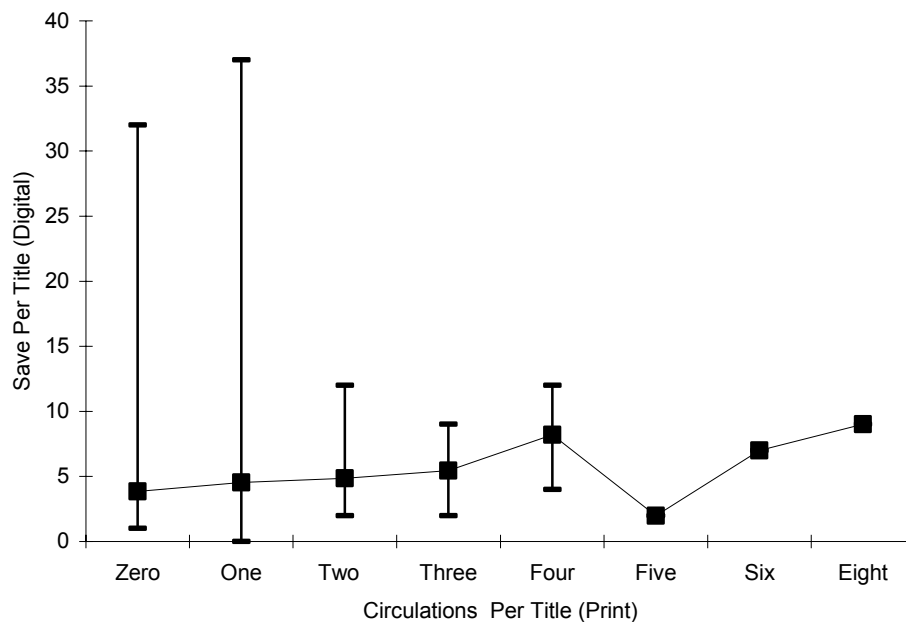
¹ The two-by-five contingency table yields a test chi-square value of 131.14 which compares to a critical value of $\chi^2(0.05,4)$ of 9.49.

² An F test of the equality of the variances of circulation for the two sets rejects the hypothesis that the variances are equal. An approximate t-test of the equality of the mean circulation has a value 13.67. We reject the null hypothesis that that mean rate of circulation of print volumes is the same with and without a full-text digital format being available too.

older titles that generally circulate less often. This evidence suggests that digital books substitute in part for printed volumes.

Analysis finds that titles that circulate more frequently in print are saved more frequently in the digital realm as well.³ This relationship is illustrated in Figure 7.

Figure 6. Digital Saves in Relation to Print Circulation (High-Mean-Low)



Reader Response - Overview

Faculty and student assessment of the project came through a series of focus groups held in 2003 and 2004. Librarians with liaison responsibilities in the humanities and social sciences recommended group participants, who were chosen for their knowledge of and experience with the collection, and for the collection's aptness to their research. Eleven Penn faculty met in the initial group in February 2003. Another group, comprised of eleven Penn graduate students, convened several weeks later. A third group, consisting

³An estimated robust regression (to adjust for heteroscedasticity) gives an estimate of 0.50 saves for each circulation in the year. Although knowledge of the rate of print circulation explains little of the variation in save activity, the slope of the relationship is statistically significantly different than zero.

of twelve individuals, included faculty and students from Bryn Mawr, Haverford, and Swarthmore, and met in February 2004.

The conversations with readers followed three lines of inquiry:

- What digital tools do you presently use in teaching, research, or course work ?
- What is your experience with the Penn/OUP/CUP project?
- What future do you see for book-length, online texts?

What the Readers Said

Use of Digital Tools

Our interviews found extremely favorable attitudes toward full-text resources from all groups. JSTOR, in particular, has deeply penetrated and positively influenced the research and study life of humanities and social science scholars. The groups were unanimous over the value of having primary sources available online. Indeed, the consensus among Penn faculty and graduate students decidedly favored primary source access over the electronic provision of secondary works. The range of primary materials cited was broad, including archival documents, authors' journals and marginalia, complete works, and rare books and manuscripts.

Other important electronic resources mentioned by faculty and graduates included digital images, especially historical maps and atlases. Graduate students at Penn make heavy use of reference and data resources online, for example dictionaries and statistical abstracts, and also newspapers.

The groups cited as advantages of full text resources ubiquity of access, the ability to overcome physical barriers, unlimited availability of texts, ease of fact checking, the ability to scan, search, and manipulate cited material, and the capacity to integrate content with learning management systems. Disadvantages included the lack of standards for displaying non-Roman scripts, the inability or expense of securing robust network connections within classrooms or at home, and the negative impact of most computer monitors on text and image resolution. But the greatest disadvantage cited is the

impracticality of reading online; the ergonomics and portability of the printed book have abiding importance to faculty and students alike.

The Project Collection

Penn faculty and graduate students were far and away the most experienced users of the digital book collection, but all the persons interviewed expressed high praise for the quality of the collection and disappointment with its lack of depth. The service contained too few titles to reward browsing.

The interviews supported the statistical findings outlined above, that online availability impacts print use. Of the Penn faculty interviewed, roughly half had identified a title in the collection and then purchased a print version. Several Penn graduate students had borrowed print editions after scanning the book online, and at least one purchased a hardcopy edition based on prior exposure to the digital book. At least one faculty from the TRI-CO colleges had opted to download several books to his personal computer in lieu of hardcopy purchase. The Bryn Mawr conversation showed interest by undergraduates in downloading parts of books instead of purchasing them. They viewed the digital library as, in part, a substitute for print. Yet one student spoke pointedly about the value of a book as object and the library as a place to browse.

Faculty and graduate students alike used the digital collection primarily to scan and evaluate content. One Penn student made a practice of using the electronic and print versions together, and viewed them as complementary. There was general agreement that while readers seek out materials digitally, they prefer to read offline. In the estimation of Penn graduate students who teach, undergraduates are exceptions to this rule. The teaching assistants in our interviews witness a greater willingness among their students to work with texts extensively online and perceive a generation difference in this area of reader preference.

There is general acceptance of the PDF format as a means of accessing full text, however, some Penn faculty characterized PDF as a conservative choice. It replicates page images

faithfully but gives relatively poor access to tables, illustrations, and footnotes. The faculty critical of PDF contrasted it to texts encoded in XML, such as the electronic books by the Johns Hopkins University Press and the History E-Book Project of the American Council of Learned Societies. More than one Penn faculty member felt the verisimilitude of the PDF was a less important feature than good search functionality, easy navigation, and image rendering. Others in the faculty group wanted select/cut/paste functions, multiple page displays, bookmarking, and TOC displays that are continuous with navigation through texts. Much of this functionality is in fact available in the PDF collection, but requires proficiency with the Acrobat software and underscores the burden of learning that interface. Graduate students had a more positive attitude toward using PDF and saw it as a superior option to presenting texts in plain HTML. One student thought it a disadvantage to have all-or-nothing downloading; it would be better if one could also save a page or individual chapter.

Penn faculty offered their perceptions of the academic impact of the digital collection. Several agreed that student bibliographies, so far, were not changed by access to books online. According to faculty, the practice of sharing URLs and bookmarks with students is a superior means of pointing them to useful resources. Faculty also saw the digital books as better vehicles for satisfying student demand for works used in courses than the library reserve room.

The University of Pennsylvania group considered the question: Would the digital library be strengthened more by adding titles or by adding features? Most agreed that adding more titles should be primary. They also responded to the question: When asked whether the Digital Books Project provides sufficient value that basic University budgets should sustain the venture once external support ends, the faculty at Penn said no without dissent. The project provides a valuable supplement to print, but not of sufficient value, so far, to justify reducing acquisitions of print books.

The Future of the Digital Book

On this issue, the interviews found consensus on several points:

- If digital book collections are to succeed in the academic sphere their content must be broad and deep. Primary works and comprehensive archives of material have the greatest value to scholars. For example, the encyclopedic works of 19th century scholarship would be highly sought after by certain classes of scholars. One Penn faculty urged publishers to consider digital compilations of the current thought in an area based on a range of literary formats, including handbooks, journal articles, and article anthologies.
- Scholars are finding images of increasing importance to their work, and with that keenly desire the capability to search effectively within texts and databases and across collections to find and extract image-based material. Advances in this area must be paralleled by advances in display technology.
- Scholars would like to see the close integration of text and non-textual information, including recorded sound, video, and data. Such integration would result in a new communication medium, one that replaces the current practice of merely rendering a book into screen-viewable form.
- Print-on-demand services in conjunction with online access would find a receptive audience among students, who tend more than faculty to rely on a certain corpus of heavily-used material.
- The digital arena may offer publishers and libraries important economic incentives and provide their audiences with unprecedented information access. However, the benefits of digital publication for younger academic authors and their fields of study are harder to discern. Our interviews found concern among faculty, and especially younger faculty, that books born digital carry less weight in promotion and tenure decisions than books published in print. This perception may hamper the ability of Presses to attract quality content for low cost, high

distribution digital collections. While digital production has the potential to resolve economic restraints on the flow of new academic books—particularly first books—it will not address the continued need for talented people to polish a manuscript, market, and review it.

Summary on Costs and Benefits

To summarize, we find that subject-based e-book collections can be quickly and inexpensively developed when undertaken as an adjunct to present-day book composition and printing processes. If the current project were scaled to 1,000 titles and licensed by just ten libraries, per title production would be approximately \$15, based on the costs incurred by Penn. This compares favorably to an average unit cost of print, which at Penn was nearly \$29 in 2003. As a supplement to print titles that receive heavy use, the digital collection model would prove cost efficient.

To optimize production costs, a press must establish the digital conversion process into its workflow and standardize, maintain, and preserve its pre-press files for use in digitization. Such workflow modifications were not in place during the early phases of the study. This had ramifications for processing time and development of the collection. As a result, one of the anticipated benefits of the digital book that did not materialize was much earlier availability than print. The cost projections mentioned above and the concerns about workflow voiced here are predicated on the use of PDF as the means for building and distributing books for online use. The issue of digital book formats is discussed further in the Conclusion.

Reader input shows, and data analysis confirms, that digital use is cursory. Scholars tend to scan tables of contents and indexes, check facts and key references online in brief exploratory sessions. A book that warrants further study is downloaded, usually for printing purposes. The download may trigger a purchase, or typically for students, a trip to the library stacks.

Virtual circulation, measured as complete e-book downloads, is intense by print books standards. On average, the ratio of complete downloads to hardcopy circulation was roughly 20:1. The small size of the e-book collection tended to inhibit the search for statistical relationships between the two formats. However, within the brief time frame and collection limits of this study, online availability did appear to reduce demand for project texts in print editions. Print editions that tended to circulate heavily also tended to get more intensive use online than books with lesser circulation.

Readers welcome the addition of digital books to other resources of the Penn Library. While small, the collection appeared to have a beneficial impact on graduate students and on certain faculty. Penn's Jonathan Steinberg, the Walter H. Annenberg Professor of Modern European History, said the collection made possible a line of current research that would have required him years to complete with only print collections. That said, all the groups interviewed were unanimous in stating the value of primary sources over secondary materials. In addition, we note that among the more sophisticated researchers interviewed, the preference is for database-style access to large bodies of text, rather than individually rendered digital facsimiles. Students, however, take a more positive view of the PDF format. Faculty interviewed said they would need more time to determine if the digital book collection, or electronic resources generally, had a positive impact on student academic performance.

Conclusion

For humanities scholars, the printed book retains its place as a fundamental resource for teaching, research, and the communication of knowledge. That said, faculty and students in this area of scholarship are growing increasingly sophisticated about electronic information. Although electronic tools, including the digital book, will not soon displace printed texts, they have become essential supplements to hardcopy resources.

The digital book is found to have particular promise as a tool that aids discovery, fosters productivity, and enables the integration of disparately located and formatted

information. For librarians, book-length full text resources offer new approaches to disseminating information and potential offsets for static and declining information budgets. Further collaboration between presses and libraries would advance the development of academically-oriented digital books and hasten the day that academics and librarians can enjoy their full benefits. The present collaboration supports a widening consensus about digital book development.

- Scanned vs. Encoded Text: PDF is well suited to cost efficient, rapid collection building. It works well for quick review and reference, as this study bears out. It has potential as an archiving medium for gray literature. It is also a potentially useful tool for constructing institutional repositories of scholarly output, because researchers who generate repository content require a modicum of skill and training to create and work with PDF applications.

However, the format is not an optimal choice to advance the more complex evolution of electronic texts. It is not easily navigated or searched, does not display search results optimally, nor is it easy to integrate with externally situated e-content, such as images, data files, or the full text of cited references. As we learned from readers in this study, scholars, even in the humanities, work today with a range of electronic resources that do provide nimble discovery functions and increasingly deeper cross-platform linkages. A more dynamic conception of the digital book, one that incorporates the presentation of full text with the robust functionality of databases and web services, will be best achieved with XML related technologies.

- Digital Architecture: The use and acceptance of digital collections will hinge on success in three important areas of development, each having a different relation to the infrastructure of future digital libraries. They are, the development of:
 - search capabilities that enable scholars to discover content stored in digital repositories and to search efficiently both within and across repositories

- linking capabilities that connect references and cited content across repositories, and
- processes that migrate content from obsolescent to emerging systems in order to preserve the intellectual legacy represented in digital collections.

With regard to discovery, large scale catalogs of full text based on broadly accepted interoperability standards, like those put forward by the Open Archives Initiative, will be critical to digital works reaching wide audiences. The use of OpenURLS and development of registries of Digital Object Identifiers (DOIs) will be essential to external linking between digital book collections and other licensed content offered by the library. An effective preservation strategy will require adoption of standardized or well-documented schema or Document Type Definitions (DTDs) applied to digital book collections. In the case of collections comprised of PDF files, adoption of PDF/A will address the preservation need.

Faculty and librarians share concerns about the impact of current economic trends on scholarly communication. As libraries redirect dollars from book purchases to journal subscriptions and database licenses, presses are increasingly constrained to publish fewer and only better-selling titles. One consequence of this dilemma is a narrowing of publication opportunities for younger faculty. With careers tied to the appearance of an all important first monograph and publication opportunities dwindling, faculty are concerned that much quality research will not see the light of day and fields of study will contract. Librarians see their own budgetary constraints along with the decline of book profits resulting in evermore homogeneous, less distinctive collections for future scholars.

In the ideal, digital distribution will eventually provide cost models that enable production of good quality albeit less profitable books. For the moment, faculty represented in this study, including one who published a digital book under the auspices of the American Historical Association, see evidence that books born digital do not yet receive the same level of attention given printed monographs in peer review and tenure

decisions. For the digital book to positively impact this quandary of scholarly communications, the bias in favor of print would first have to ease.

Although the evolution of scholarly publication and the technologies serving it are still in flux, the present study demonstrated that digital books can work effectively as a library of research materials. Even within the humanities, where the transition to electronic information has been regarded as occurring at a slower pace than in science, technology and medicine, large collections of book-length texts online will find an enthusiastic audience and receive heavy use.

Oxford's Response

Niko Pfund, Vice President and Publisher, Academic, Professional and Medical Books for Oxford University Press, provided these insights from the OUP.

Oxford has a number of digital publishing initiatives underway and, although these have all been informed by numerous focus groups, market research, and other forms of solicited feedback from librarians and scholars, the conclusions that emerge from the report are extremely useful to us, both confirming and refuting certain assumptions.

Perhaps the most positive development hinted at herein is the apparent embrace by scholars in the humanities of online resources. The scientific academic community was clearly first out of the gate in the adoption of online information, and the more gradual migration of humanists and social scientists to the web has been, and continues to be, a major factor in the Press's thinking and planning.

The study's conclusion that a digital book is often treated as a supplement to a printed copy, or as a means of finding or previewing content in the printed book, is also encouraging. There is also intriguing statistical evidence that digital books substitute in part for printed volumes. This issue of "substitution or supplement" is a central one for publishers and deserves further study.

The apparent preference for robust coding with enhanced search capability over PDF is of particular interest as we continue to weigh the added costs of such coding vs. the benefits. The question of cost vs. benefit looms large in the choice of XML over PDF.

The unanimously expressed desire of students and scholars for online access to primary, rather than secondary sources was striking. This distinction has been the subject of considerable internal debate, especially as the Press invariably compares any prospective digital publishing initiative to extant ones such as JSTOR.

One of the most bedeviling aspects of creating an online collection of book content, that of third party rights clearance, was moot given that the Digital Book Project was never intended as a commercial enterprise. The related fact, that users seemed particularly to value digital images, especially maps and atlases, is good news for ARTSTOR and suggests that publishers would be well-advised in the future to create their own such images, rather than license extant images from other copyright holders. This would allow for greater flexibility with regard to online distribution.

Another topic of considerable internal discussion at Oxford is the logistical and market-oriented pros and cons of a single-publisher archive vs. a multi-publisher one. The logistical obstacles of the multi-publisher model are apparent in the report's allusions to the technical and coordination problems involved when seeking to extract materials from numerous production sites around the world. In fact, the report astutely covers many of the technical challenges involved in any digitization project, particularly standardization problems often requiring expensive manual file manipulation.

The following points confirm what is known, but are no less valuable therefore: that collections should be broad and deep, that images and non-textual materials are desirable, and that a Print on Demand function to supplement online access is desirable. The point about Print on Demand in particular seems to many in the publishing business a logical endpoint to database publishing

Perhaps most encouraging was the broad distribution of saves across titles, confirming the conclusion of the BYTES (Books You Teach Every Semester) study of several years ago that university press content is consistently relevant and frequently consulted.

Finally, what really stands out is the finding that digital use is cursory, suggesting that online research exists as an add-on to a deeper, immersive read, rather than a possible replacement. Of particular concern is the possible trend, expressed with some alarm in several recent reports, that readers are skimming more and reading less, that intellectual highlighting is perhaps not simply a function of online reading but is becoming a procedural shortcut to deeper learning, regardless of format. This raises again the above-stated interest in a more conclusive treatment of the "substitution or supplement" question, which may only be possible with a larger sample than the one available here.