

The Columbia University Evaluation Study of Online Book Use: 1995 - 1999

Paul B. Kantor, Tantalus Inc. and Rutgers University (*kantor@scils.rutgers.edu*)
Mary Summerfield, Consultant, (*marysummerfield@earthlink.net*)
Carol Mandel, New York University, (*carol.mandel@nyu.edu*)

1 Introduction

This paper is a partial report on some observations about cost, use, and users of online books during the Columbia experiment. The project began in 1995, ended in fall of 1999, and has been supported by The Andrew W. Mellon Foundation. The experience involved integration of two very diverse cultures, and has taught us the relevance of the following joke.

A manager, an engineer and a computer scientist are all traveling in a car in the mountains when the brakes fail and the car careens down the road and eventually stops just hanging over the edge of a cliff. They carefully climb out of the car and the manager says, "Well, now we'll have to form a focus team for a matrix review of vision and objectives." The engineer says, "Let me have a screw driver; I may be able to fix this in 10 minutes". And the computer scientist says, "Let's push this back up to the top of the hill and see if the brakes fail again."

Our approach to online books at Columbia was like that of the engineer, but "10 minutes" has been more like four years. One of "lessons learned" is that, as libraries become more interdependent with computers, we must become more accustomed to the kind of trial and error approach exhibited in the joke.

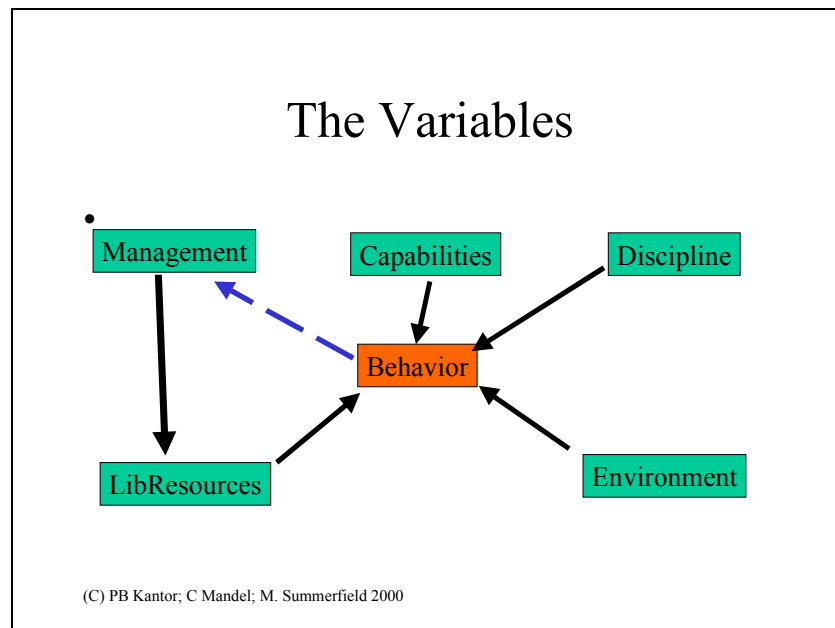


Figure 1. Abstract Variables of Interest

We started from an abstract formulation of the relation among variables, as shown

in Figure 1. Our goal is to understand the behavior of the user of the system, shown in the middle of the diagram. The capabilities of the individual users obviously influence their behavior. Their disciplines also probably influenced it. The overall environment including technology and attitudes toward computers influence it. And, of course, the resources available in the library influence behavior. In turn, library management controls those resources.. From one perspective, the study reported here is an effort to insert the blue line shown in Figure 1, to provide management with feedback about the behavior of the users which it can use to better manage the library resources.

The project began in early 1995 and ended in late 1999. The project prepared books in HTML format. This choice seemed reasonable at the time it was made (1995), although later observation of user behavior makes us less certain of that choice. The evaluation component of the project included monitoring of the national technological environment. Overall we tried to take an “economic perspective” on the entire complex issue of establishing an online books component at an existing major research library¹.

2 Economic Perspective

In taking an economic perspective, we presume that all the actors in the situation weigh the costs and benefits of various alternatives available to them. Each applies some kind of personal utility function to those costs and benefits, and chooses the action with the largest personal utility. In this complex setting there are many different kinds of “economic persons”: student persons, faculty persons, and staff persons. And in an economic sense, the library itself and even the entire university can be thought of as “persons” (see Figure 2).

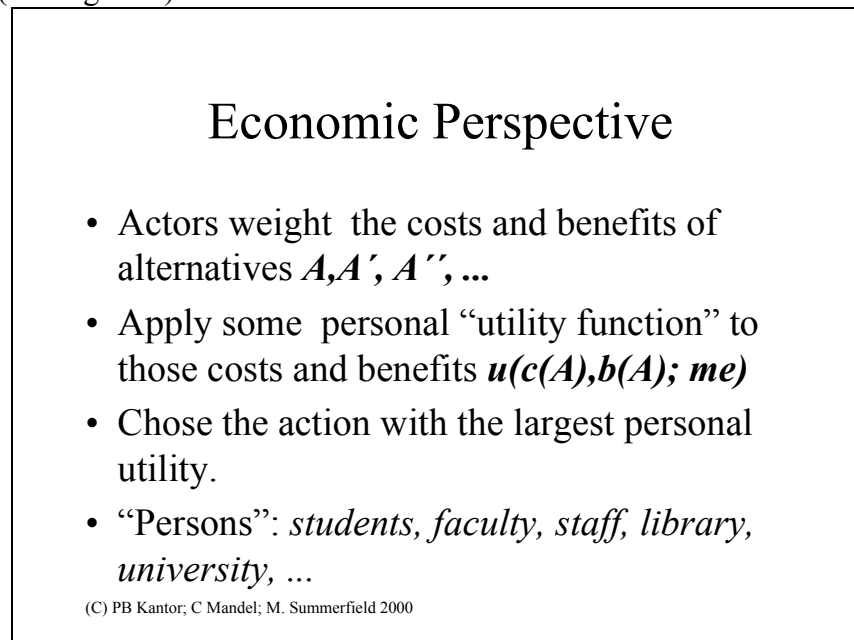


Figure 2. Economic Model based on Utility

2.1 Individual economic factors

2.1.1 User Costs

What are some of the forces affecting individuals? First, there are costs of two kinds. There are capital costs: one is the cost of equipment needed to be able to use the digital library or online books and the other is the cost of acquiring the skills that are needed. There are also continuing costs. Since, in the setting of our project, there is no transfer of funds from users to the library associated with use events, those costs are really (a) the cost of connecting to the library and (b) the mental costs or efforts associated with use. Not a lot is known about these costs to the user, at this point. But we have some feeling that in the transition from page-based books to the HTML format that we chose, certain kinds of mental landmarks that readers have developed over years of working with print on paper are removed and it seems likely that this results in additional mental cost to the users.

2.1.2 User Benefits

There are also benefits to the users. First among these, of course, is ubiquity of access. In addition, the book-marking system (supported through the browser) permits them to store pointers to important locations within an extended text. Our system did not directly support annotations, but obviously annotations can be established in the users' own computers. Our system provided a search capability. And, of course, using a system like this provides the intangible benefit of being up-to-date relative to one's peers.

Beyond all this, having and using a digital library provides "symbolic utility." Symbolic utility is a concept introduced by the philosopher Robert Nozick to represent the utility assigned to something that it is good to have or to do, even if it doesn't necessarily "work". In the book where he introduces the idea (Nozick, 1993), Nozick cites "prayer" as an example of an activity which, in the opinion of many people, is worth doing even if it doesn't immediately deliver results.

To sum up, there are a variety of benefits and it seems most probable that they outweigh the costs to individual users.

2.2 Staff Economic Factors

2.2.1 Staff Costs

There are also forces that affect the staff of any library that introduces a substantial digital component. There are two important kinds of costs. One is a kind of learning curve, representing costs that must be incurred in order to get the system to work. The other, which is becoming a pervasive feature of the library world today, is the cost of continuous change. Working in an environment of continuous change introduces psychological stresses. Many older librarians do not enjoy those stresses and the profession did not select for people who do enjoy them. Thus we will probably see a gradual change in psychological profile of the profession as it becomes one in which continuous change is a fact of life.

2.2.2 Staff benefits

Among benefits to the staff, the first and most important is the ability to provide better service to patrons. Another important benefit is the ability, in the online books or digital library situation, to adapt materials developed by others. In focus groups conducted at New York University, one of us heard for the first time librarians reporting that they were pleased to be able to develop web resources in which pointers to resources developed by librarians at other institutions played a major role. [Kantor, pc.] A final benefit to staff is the fact that by working in the digital environment they are developing skills that are much more portable than traditional library skills. We must anticipate that, in the future, some fraction of a university library's staff will move to the "dot-com" electronic commerce setting expecting to promptly earn millions of dollars. As a practical matter, we want to treat them very well, so that they can later endow the digital equivalent of a new reading room.

2.3 *Library Economic Factors*

2.3.1 Library Costs

Turning to the forces affecting the library as a whole, the costs are obvious. Uppermost are costs of equipment, costs in the development of materials, and costs associated with training.

2.3.2 Library Benefits

Among the important benefits are the contributions to the competitiveness of the university, and the contribution that "making the library digital" contributes to the shared professional goals of growth and service.

2.4 *Publisher Economic Factors*

The publishers must consider the potential of electronic books in terms of their business plans and goals. Publishers are allies of the libraries, of the authors, and of the readers. Yet the relationship is sometimes conceived as an antagonistic one, because some portion of the price to readers and libraries goes to the publishers rather than to the authors. In other components of this study we have examined publishers' costs, but we do not report on them here. We presume that for-profit publishers seek to maximize profit, while non-profit publishers seek to maximize the net of income over expenses attributable to each book.

3 The Studies

Sections 2.1 to 2.3 above constitute the general economic framework for the user/use components of the Columbia University Online Books Evaluation Project. Having laid out the assumptions described above, we undertook to study the environment, publishing costs, and library costs. We explored various views on the functions and design of online books. We conducted numerous and diverse studies of use and of user preferences. This paper summarizes some of what we've learned and discusses implications for the future.

When viewed in concrete, rather than abstract terms, what the Columbia University Online Books Evaluation Project did was repackage books for online delivery, study the use of those books, and estimate the costs for publishers and libraries of providing print and online books. There were four publishing partners in the project: Columbia University Press, Oxford University Press, Garland Publishing, and Simon and Schuster Higher Education. We analyzed the costs of development, delivery, and use and we sought to relate those costs and that use to the context and to the potential for service. The analysis of the potential for service is by no means complete and we would rather say that we have just begun.

3.1 Why put books online?

To review some familiar points, why should we put books online? The first point is that we anticipate that online books will be cheaper to produce, to purchase, to acquire, and to maintain. We expect that online books will provide increased functionality such as searching and linking. They offer obvious potential for enriched content through the addition or linking to multimedia, computer simulations, and other features. There is also potential for developing expanded products, which are something more than a single book, and rather like a collection of books linked through a web site. And of course, online books can provide availability around the clock and calendar.

3.2 Why not put books online?

3.2.1 Usability

We also ought to ask “why not put books online?” This was a serious issue in 1994 and 1995, when the project was planned and launched. At that time, it seemed that the most important negative point was that “no one wants to read books online”. We do not know how true this is in the year 2000. There are definitely many who do not want to *read* books online, but we must entertain the possibility that most of them are as old as the scholars and librarians at this conference, and that they will eventually be replaced by people who *do* want to read online. When the project began, it was anticipated that online books would be difficult to use. At that time (1995), it was not even apparent that web technology would be easy to use. We were also concerned that there was no feasible market model for the development of online books. We cannot say today that there is a clearly defined market model, but at least the activities of netLibrary, Questia (and others) show that there are multiple possible paths into the market for scholarly books, aimed at libraries and students, respectively.

3.2.2 Accessibility

We were further concerned about the adequacy of access and connectivity. We had in mind, primarily, people working at home, connecting over telephone lines with top speed of 14.4 kbps, which seemed likely to be inadequate. It is interesting to note that about halfway through the project the typical home-access speed had moved up to about 56 kilobits per second, and it has not advanced much to this date (2000).

3.2.3 Production Cost

We were concerned that online books would be too costly to produce. In fact, we shall see that when they are produced in the way that we have done, they are very costly.

Although, when compared with the total life cycle cost of paper, they are still something of a bargain. As we will hear during this conference, publishers are working hard to reduce those prices to make online books very competitive. We also had the feeling that authors might oppose the presentation of their books in online form. Factors include feared loss of royalties, and fear that rendering a book in HTML would remove some important aspect of layout and feel that was important to the author. Most importantly young scholars might fear that exclusive publication in an online form might become common for first time authors and would demean their works and lessen their chances for career advancement. Alternatively, however, established academic authors, at least, are very concerned with being able to document the *impact* of their works and the extent to which they are being read. The online environment is ideal for this.

3.3 National Environment – Access

Reviewing changes in the environment for online books from 1995 to 1999, we see a number of factors. First, of course is the improved price to power ratio for personal computers, discussed further below. We saw penetration of Internet use to more than 50% of all USA households by 1999. In 1999 half of all adults in the USA were Internet users. On the other hand, as noted above, there has been improvement in neither speed nor Internet service provider pricing since 1997. 1998 did see the emergence of hand-held book readers, and some of our focus group work suggests that this will be important in the growth of the online book market. (Fig 3. about here)

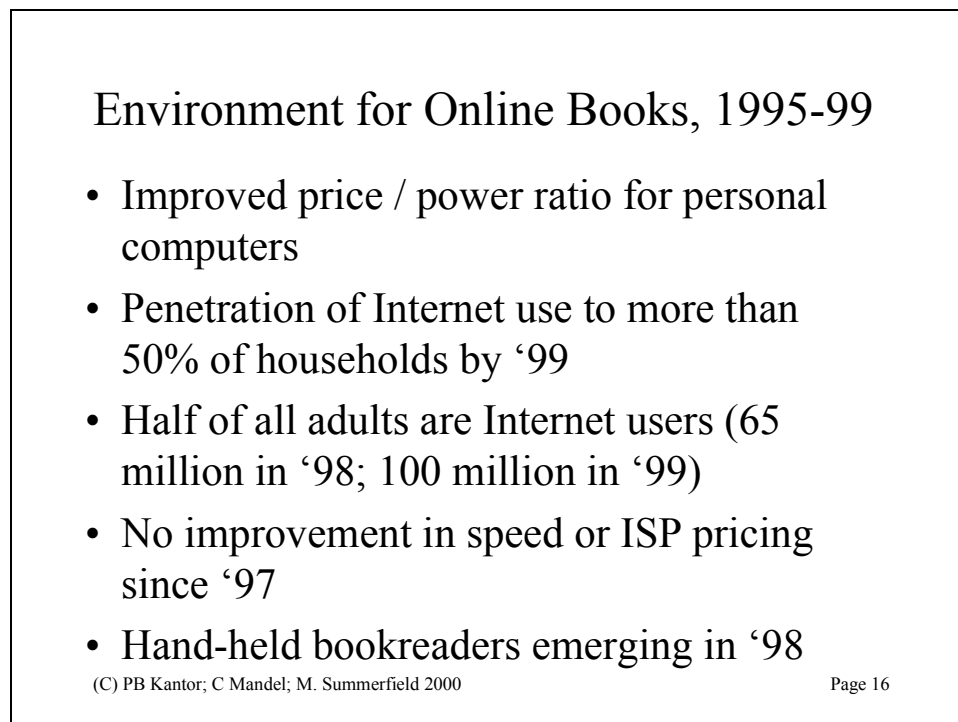


Figure 3. Evolution of the Technology Environment

3.4 National Environment - Computer power

As shown in Figure 4, Moore's famous law (that technology doubles in power every 18 months) may be true, but the expected corollary, an exponential drop in costs, is not true. Starting from a time when the base price of an adequate computer was about \$4000 we would have expected that by the end of our study this price would have dropped to well below \$1000. What we actually saw, through a program of tracing ads for computers, is that prices dropped fairly rapidly to around \$2000, which seemed to be a kind of sticking point, and they held at \$2000 for some time. There has just recently been a new break down to \$1000. It seems clear that the strategy of manufacturers is to identify market points that seem acceptable and to increase the strength of the computers rather than drop the price past those points. We estimate that if Moore's law held strictly, a general purpose computer adequate for the use of online books over the 56 kbps lines should now be only something like \$300.

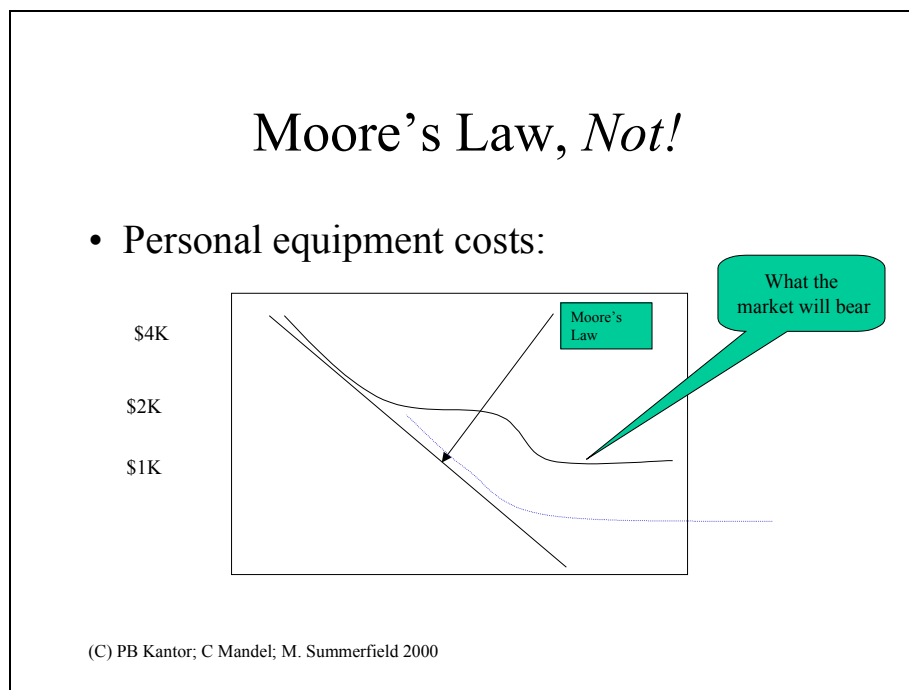


Figure 4. Deviation from Moores's Law

3.5 Local Columbia Environment

The local environment at Columbia for online books changed substantially during the period 1995 to 1999. By the end of this period there was Ethernet connectivity to every building and dormitory. By 1997, which is the last time that we could justify the costs of surveying to ask the question, 80% of students and faculty had adequate access to a network computer. By 1997, most library users reported an average of six hours per week of online activity of all kinds. That works out to about an hour a day and we estimate that by now this has probably at least doubled if averaged over the entire community.

By spring 1999, online full text use had become common at Columbia. For example, the monthly level of JSTOR use was equal to the entire population. (Note, this does not suggest that every single person used it, but rather that the number of uses had become equal to the population). We found that most online book use was from on-campus computers. In other words, our original concern that access from home might not be adequate has not been allayed, but it seems people find other ways to get to the books. It is quite possible that, as bandwidth to the home increases, the usage of online books will increase further.

4 Cost and Use data

We developed a variety of sources of data in the online books evaluation project, including surveys: online, mail, by telephone and in class. We also conducted individual in-person and telephone interviews of scholars and a number of focus groups involving users, potential users, and librarians. In this report, we focus on cost analyses and on web data. The reader may visit the project web site [www.columbia.edu/cu/libraries/digital/texts/about.htm] to review other studies and reports.

4.1 Production Costs

In a print production environment, online is an additional cost and we found an amazing range, from four cents per page to more than \$2.00/page, which works out to a range of something like \$100 to \$1000 per title. The range of cost is due to the enormous variations in publisher's tapes as they are available at this time and in the conversion process employed by various projects. Achieving the low-end cost requires a very standard and well-behaved PostScript electronic file and PDF output. In addition, these figures include some unknown component of experimentation cost, as this project and others adapted to variations in input, and in desired presentation format.

Sample e-book production costs

- Conversion: OCR, SGML • \$ 1.51 / pg.
- Conversion: ASCII to HTML • \$ 1.00 / pg.
- Conversion: PS to PDF – \$.04 / pg.
- Conversion management • \$20 / title
- Books on server • \$ 1 / MB / yr.

(C) PB Kantor; C Mandel; M. Summerfield 2000

Page 23

Figure 5. Conversion Cost examples

In Figure 5 we present some sample electronic book production costs. One conversion route is from OCR (or from SGML), and the other, somewhat less expensive, begins with ASCII and going to HTML. Conversion from PostScript to PDF is done using software from ADOBE and yields a cost of about four cents per page.ⁱⁱ Note that this process, which has been tested at the University of Pennsylvania, does not yield fully navigable HTML files, but yields PDF output only. Management of conversion is estimated to have cost about \$20/title at Columbia. Maintaining books on the server is steadily less expensive, estimated by the end of 1999 to cost about \$1/megabyte per year. Through conversations with scholarly publishers, we have been able to estimate that the potential savings for moving to online format, without paper would be about 10% at the plant (that is changes in typesetting costs) and perhaps an additional 15% in costs avoided in paper, printing and binding. Also eliminated would be costs associated with warehousing and shipping which we did not attempt to estimate.

On the other hand, there are offsets to these savings for online production. There are, for the publishers, costs associated with marketing, and with customer service. There are costs, which may ultimately be assumed by publishers or by libraries, associated with continuing file maintenance and migration. A "rational economic" publisher will only maintain the file for a book as long as the discounted total expected future revenue from sales exceeds the total discounted projected cost of keeping the file. Thus, libraries cannot rely on publishers to maintain the files of books with very low demand, unless they are willing to pay service fees that cover the publishers' expenses.

It is always risky to present cost figures unencumbered by the detailed caveats

that surround their estimation. However, we do have an estimate that life cycle costs, to the library, for online and paper books. These, projected over a thirty-year life cycle and discounted at a 5 percent real cost of money, are lower for online books. The rough breakout is shown in Table 1. While the last two or three digits of each number are not important here, what we see is that essentially the change is equal to the avoidance of costs associated with circulation. In the long run the costs for online books would most likely be quite a bit lower as copy cataloging would prevail rather than the original cataloging experienced and included in the costs for this project. Original cataloging cost about \$25 per title while copy cataloging would cost significantly less per title.

Lifecycle (30 yr.) costs are lower for online books		
	<u>Print</u>	<u>Online</u>
<i>Acq/Proc.</i>	\$47.26	\$38.51
<i>Storage/Maint.</i>	14.34	38.43
<i>Circulation</i>	43.97	(incl. above)
Total (30 yrs.)	\$105.57	\$76.94

(C) PB Kantor; C Mandel; M. Summerfield 2000 Page 26

Table 1 Estimated Life Cycle Costs

4.2 Design Considerations - Librarians

We conducted focus groups with librarians to identify market and design features that they consider important in building a collection of online books. The first feature emerging is the ability to search across selected *groups* of titles. A second, rather technical issue is the existence of "stable, granular" URLs. Stable means, of course, that they remain the same over time, or at least that the system does not have to be manually updated. Granular has to do with the level of specificity with which a user can access a book. In the Columbia approach to online books, an individual file corresponds to a chapter within a book. We found, of course, that librarians want good bibliographic control of online books, with direct linking from the catalog into the book. But they would like to see usage data on individual titles in some standard form. This usage data can feed back to rationalize (both in the sense of providing justification, and in the sense of making more rational) online book acquisition policies. Finally, librarians want to be assured that an online book system will support reliable migration to new platforms.

4.3 Design Considerations - Scholars

Both in-depth interviews and focus groups with scholars generated a somewhat different list of desired design features. Scholars would like to be able move directly into the online book via direct link from the online catalog. They would like to be able to define groupings of texts, on the fly, and search across that collection of texts. They would like a comprehensive and detailed table of contents, with direct linking into the book (providing, in effect, analytic indexing). When images are a significant part of the text they would like to see browsable, linked, thumbnail images. They would like screens and displays supporting the ability to show two pages at once, permitting comparison of what the author said on page 82 and the apparent contradiction that showed up on page 314. They would like to be able to see the footnotes and the text in parallel displayed on the same screen, even if the "footnotes" are actually endnotes. They would also like to see pagination matching the print version, which refers not only to the need for navigational bearings, but also to the fact that frequently the citation that led them to a book specified a particular page.

Scholars would prefer that, whenever the collection contains the relevant material, references be hyperlinked directly into the cited material. They would like to be able to link to a dictionary (it was unclear whether they wanted every word to be linked to the dictionary or only the words that they personally didn't know). They would like to be able to adjust fonts and formats for easier reading on screen. They would like to have annotation and highlighting capability that they could store with the book. And, although scholars often are considered to be among the most competitive of individuals, they did express an interest in having the ability to share annotations on a single text.

5 Study of Users

The remainder of this paper discusses some of what we have learned about the users. The first interesting point is a relation among three concepts or variables: technology, behavior, and attitudes. We expected that the technology, as it grew, would influence the attitudes of scholars, both faculty and students, which in turn would influence their behavior. However, we tracked attitudes carefully over the entire study and saw only the smallest movement towards believing that online books are a better way to do one's scholarly work. This forces us to conclude that, in fact, technology effectively influences behavior and that attitudes simply have to catch up. This may mean that scholars are moved to technology by a subliminal perception of benefits, which they cannot articulate. On the other hand, it may mean that fashions in scholarly behavior are simply no more rational than any other kinds of fashion.

5.1 Tracking Individual Users

The key analytical innovation in the Columbia online books project was the introduction, in 1997, of the ability to identify the activity of unique users. This was a

fortunate byproduct of the security system, developed to permit people to read online books from home. To maintain confidentiality of the users, the actual process was that system analysts processed the data files replacing the identities of individual users with the not very informative labels, User 1, User 2 and so forth. In fact, each time that code was run, the name "User 1" was assigned to the person who had used the system the most and so forth. Thus, when we did the analysis at a later time, User 1 did not refer to the same person anymore. Fortunately, at least up to the close of the project, all data files were stored and so we were able to do a complete cumulative analysis.

5.2 Analysis of individual use Persistence and Adoption?

With anonymity thus ensured, we were permitted to link usage to administrative files containing demographic information about the users. Typical results are those shown in Table 2, reporting the distribution of the status of individual users at the time they first used a particular resource. The resource in this case was the online version of the *Oxford English Dictionary*. While we had a number of reference works available online and, by the close of the project, close to 200 books in online form, the total usage of the *OED* represented approximately 50 percent of all online usage, and so it is used here to illustrate the types of analyses that were performed.

Status At First Use of OED			
Value Label	Value	Frequency	Percent
Undergraduate Studen	2.00	2088	58.0
Other	5.00	607	16.9
Missing	99.00	328	9.1
Graduate Student	1.00	295	8.2
Other Student	3.00	145	4.0
Faculty	4.00	136	3.8
		-----	-----
Total		3599	100.0

(C) PB Kantor; C Mandel; M. Summerfield 2000

Table 2. Status of Users at Time of First Use

There were 3,600 individuals who used the *OED* during the study period. Just over 2,000 of these were undergraduate students at the time of first use. Nearly 300 were graduate students and close to 140 were faculty members.

We analyzed the ways in which individual users used the resource. To do this we

graph would be roughly linear. We can show what the actual data look like (again concentrating on the *OED*, which had heavy use) as shown in Figure 7.

Figure 7 is a scatter plot. Each point represents one individual user. The y-coordinate of the point represents the number of sessions that an individual had with the *OED* and the x-coordinate represents the number of days since that individual first used the *OED*. Some kind of very steep line like the one marked "would be nice" is what we had hoped to see. In fact, a regression analysis, which seeks the "best straight line" consistent with the data, shows that the best fit is the middle one of the three lines that run nearly horizontally across the picture. The fact that it is nearly horizontal means that there is not very much use over time by individuals. The parallel upper and lower lines represent a 95% confidence interval for the prediction, and it is apparent that many points (representing many users) do not fit this "best model".

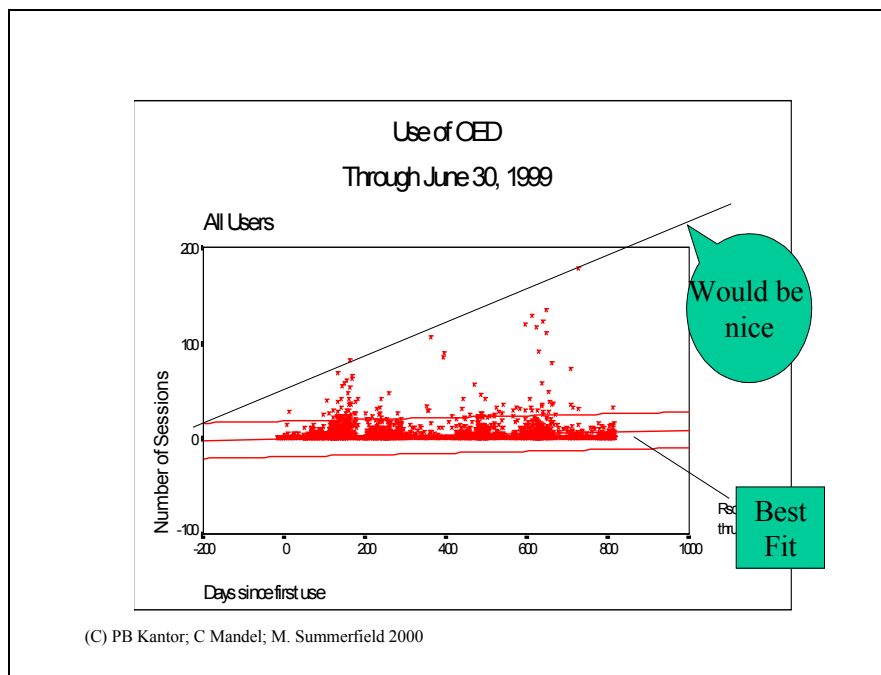


Figure 7. Scatter plot of totla use against time since first use

We can plot this data in a more familiar form by showing the distribution of time since first use, without paying attention to how much use there has been. This is the projection of the preceding figure onto the horizontal axis (Figure 8). We see, as have most researchers in the academic setting before us, that it is very easy to discover the existence of the semester. Each of the four peaks in this graph corresponds to an academic semester. There might be some cause for optimism in the fact that the left most peak, which represents the most recent surge in use, spring 1999, seems to rise higher than any of the earlier ones. However we don't know quite what to make of the fact that the one before it (fall 1998) represents a drop from the preceding fall.

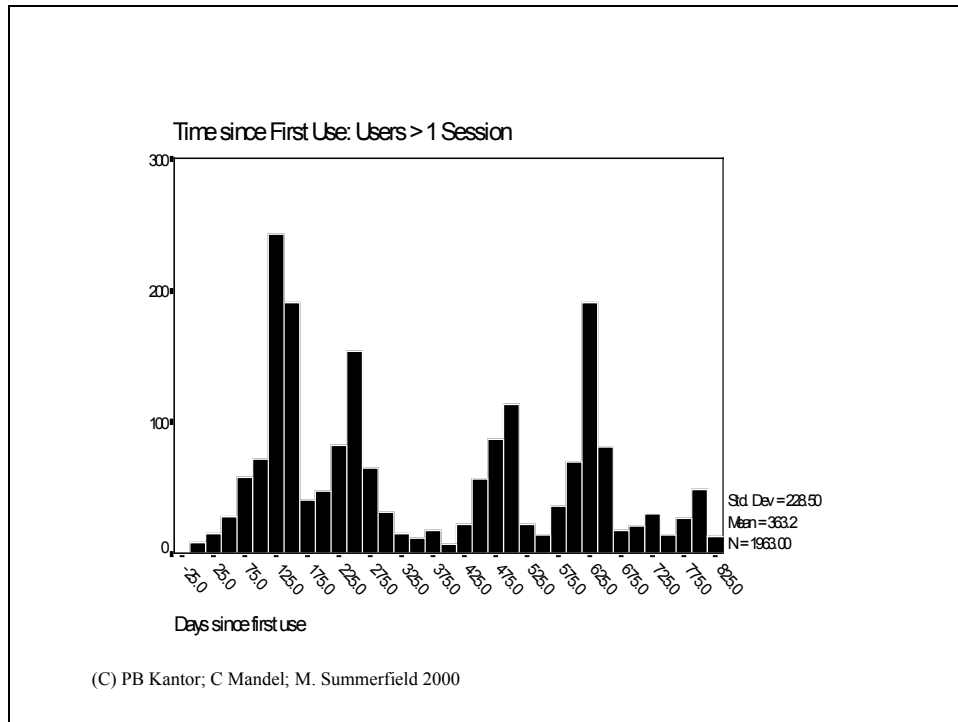


Figure 8. Histogram of Time Since First Use.

5.3 Online Versus Paper Modalities: Usage Data

Our data (based on comparison between the online book usage figures and data collected through circulation statistics and slips placed in corresponding reference titles in the library) indicate that online books were used more than their print counterparts. If we count circulation alone we find that there were about three times as many readings per book online as for the paper version. After consultation with librarians we believe that a reasonable correction for in-house use is to increase circulation by 50%. This would reduce the “3” that we observed to a ratio of 2.

We were also interested in studying how people use online books. We have approached this in two different ways. One is essentially qualitative in which we asked people in surveys and in interviews how they used online books. In doing that we were able to identify at least the following kinds of activity: browsing, grazing (that is, reading portions of text scattered through the book) citation checking, the finding of individual facts or quotations; reading on reserve for a course; determining the need for a paper copy; printing (that is, turning the online book into paper); and directly reading online.

We have also, because we can track individual users, been able to break some new ground in quantitative analysis of how people use books online. Analyzing the sequence of clicking on book sections (recall that, generally, each chapter is a separate file, and hence a separate entry in the web sever log) we are able to distinguish a number of different ways in which individuals use online books. The first way is purely linear. In truly linear use, an individual reads chapters of a book in exactly the same order in which they appear in the printed volume. The second pattern of use is quasi-linear, in which the sections of the book are visited in some personalized order but each section is read once and only once. Then, of course, there is the truly hyper-linear use of the book

in which sections are visited in an arbitrary order and some sections are visited more than once. But this occurs only about 12% of the time. Our observations indicate that most use of online books is still quasi-linear (with index used only at the start) or perhaps “indo-linear” (multiple index uses).

Patterns in online use (Quantitative)

- Linear: A,B,C,D,
- QuasiLinear: D,C,E,G (each only once)
- Uses of index:
 - I, D,C,E,G
 - I,A,I,C,I,G ... “Indo-Linear”?
- Hyperlinear: A,G,S,A,T,G,...
- Most use still Line, QL or IL

(C) PB Kantor; C Mandel; M. Summerfield 2000

Table 3 Patterns of Online Use

There are several ways that a use pattern may involve use of the index (or, more generally, search tools). The first format is to use a search tool once, at the outset, and then to view portions of the book in some linear or quasi-linear order. Another possibility involves using the index, going to a section, and then going back to the index and out to another section and continuing in this pattern. We might call this “indo-linear”. Whether this is a natural behavior evolving in the presence of online books or an

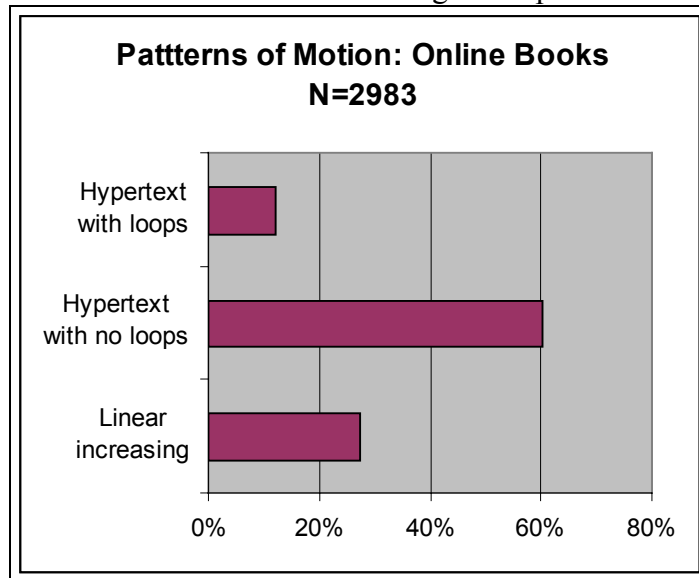


Figure 9 Patterns of Motion in Online Books

artifact introduced by the fact that returning to some index or search tool may be the easiest way to get to the next section is something we don't know at this point. In thinking about these patterns of use, we may compare them to what a person might do with the book in hand, at the library shelf, or with access to the catalog, in some online format. We must recall that having purchased a paper copy for the library does not ensure that the book is available. The book might be in circulation, or the book might be mysteriously missing and so not on the shelf. The book might be in the library where it should be but if the library is closed the book is not available to a user. And finally we note that having the book in the online public access catalog does not support even the roughest form of browsing into the book until the book itself is put online. The catalog provides so little information about any book that a scholar might not be aware that one contains material relevant to his work. If so, the mere ownership of that book by his library does not make it truly available to him. Catalog records enhanced with tables of contents and book indexes are a relatively new offering and a major assist to the scholar in locating books relevant to his or her research.

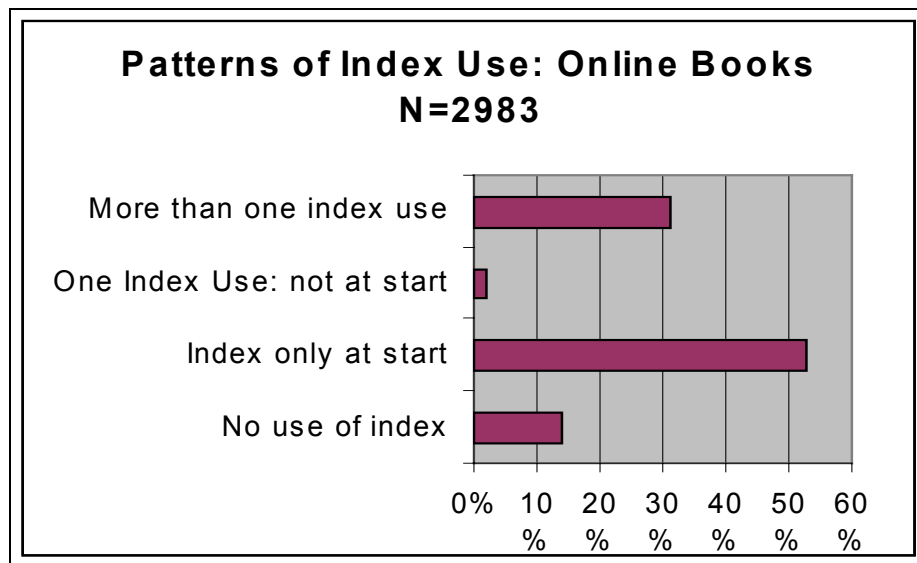


Figure 10 Use of Index in Online Books

Hence, the online access to a full book represents a quantum leap in the availability of the contents of that book, and, we believe, lowers the barriers to access for many modalities. Perhaps the only modality for which it is not clear that online access is preferable is “plain old reading at length”.

6 Economic Models

6.1 Economic models for the Behavior of Scholars

Given our original framework, we would like to bring together everything that we have learned, to formulate some economic model about scholars' preferences for modalities of book access. We believe that, for this issue, one key variable is “how much

it costs” which we characterize simply as low or high. (For the moment let us imagine that this is the purchase price of the book, as far as the scholar is concerned.). We propose that the other key variable is whether the scholar intends to “read much” or “read little”. In saying this we gloss over a complexity. There is only one way to read a little of a book: read a little of it, at one time. But there are two ways to read much of a book: read a great deal of it over a short period of time, or refer to it many times over a long period.

We believe that whether the book is cheap or expensive, if only a little of it is to be read, the scholar will prefer to get it online. Based on data available to us during the span of this project, we believe that if “much” of the book is to be read, the scholar will prefer to get it in paper form. If the cost is low, the scholar will buy it; and if the cost is high, the scholar would like the library to buy it so that he or she can borrow it.

Scholars’ Preferences for Book Access		
	<i>Read Much</i>	<i>Read Little</i>
<i>Low cost book</i>	Buy	Online
<i>High cost book</i>	Borrow	Online

(C) PB Kantor; C Mandel; M. Summerfield 2000 Page 41

Table 4 Factors Influencing Access Preferences

In effect, what we seem to find is that users want online books for convenient access and for assured 7 days by 24 hours per day availability. And they want online books for many of the purposes listed above in the qualitative descriptions. They are particularly attracted by added functionality of annotating and hyperlinking. An example which was not implemented in the Columbia model, but which shows one aspect of the future is the NEC CiteSeer model. CiteSeer has processed postscript versions of technical papers and, for any given paper, the reader can access passages from other papers in which that particular paper is cited. This gives rich contextual information about what the paper says and how it has been seen or used by other scholars. Nonetheless, the results of our studies seem to show that when scholars want to read books at length, they still want them in paper form.

6.2 Economic Perspective of Librarians

Complementary to this analysis of when scholars will prefer online books, our

focus group studies with librarians indicate that librarians want online books for high demand books (for example instead of buying a second copy). Librarians also want online books to meet transient demand, rather than having to purchase additional copies which will be unused later. And, of course, librarians want online books for the anticipated cost savings.

On the other hand librarians are concerned about paying twice (that is having to pay separately for the online version of a book that they hold in paper). They are concerned about the uncertainty of preservation and migration of digital forms. And they are particularly concerned about the appearance of unwanted and unused material in bundled packages. While bundling in general can increase both consumer and producer benefit (Shapiro and Varian, 1999), librarians are particularly concerned with the flow of cash from the institution to the publishers, and they would like to have the finest possible detailed control to optimize the allocation of those funds, by avoiding materials that are less in demand.

6.3 Speculations on Market Models

We have tried to speculate on options for library-oriented models for the introduction of online books. For example, one might imagine that online versions are made available for little or no additional cost to purchasers of paper copies. One might hope to see entire collections of online material priced very attractively. On the other hand, opposite to that bundling, one might see some kind of on-demand licensing, or on-demand print ordering. One market entrant, netLibrary, is replicating the system of selling print books in providing online books to individual libraries or library consortia and allowing just one user at a time (a circulation) for each of those books. On the other hand there are economic models that might be said to be more consumer oriented and less tailored to the concerns of a library. These include Questia's effort to build an online book collection the size of a college library (250,000 volumes) and to sell subscriptions to students. Another path is the hand-held device and downloadable book that is now coming to market. Generally speaking in a consumer-oriented model, pricing of the electronic form will be unrelated to print purchase, as there is little chance that consumers can be persuaded to buy the same "book" twice.

We speculate, but at this point can only ask, whether different models will emerge for different classes of print materials such as text books, scholarly books, and narrow interest (sometimes called endangered) scholarly books.

As we reach the end of our study period it appears that a number of transitional compromise models are available or being developed. The leading one is the provision by publishers of print and online publication. Among other virtues of the model, there is the possibility of electronic publication of both a backlist (that is the books that have been available for over a year) and a front list (that is the books newly published). Since publishers still need to protect ultimate paper sales, there may be some kind of limits on functionality that would be provided for new titles that are presented in front list form.

Online books have some implications for knowledge generation. Preparation of certain kinds of scholarship, involving concordance, or detailed textual comparison, can now be semi-automated. Use of online books can be tracked at a micro level, providing valuable information for authors and publishers. In fact, scholarly authors must become concerned about these data since their advancement may depend on being able to document the degree to which their works are used, as well as the degree to which they

are cited.

7 Concluding Unscientific Postscript

Having studied the whole situation for four years we feel emboldened to make a few predictions, although we don't guarantee that they will be correct. It appears that complex functionality will be reserved for books that have large sales or are developed in subsidized projects. We anticipate that endangered monographs will be available from academic or society servers, from sites like the Los Alamos Preprint site (<http://xxx.lanl.gov>), or from the individual authors themselves. In other words, they won't be "published" as we understand it today. Many books will appear in both electronic and print versions. Commercial enterprises or academic organizations and not library experiments will define the product that eventually comes to dominate.

In a more light-hearted vein, there are a few predictions that we can make with confidence. It appears, at least in the short-term future, that no one will save money. It is quite likely that someone will make money, and perhaps, in a sort of "dot com" revolution, will make a great deal of it. To sum up we may say that civilization as we know it will be either (1) Transformed beyond recognition; (2) Essentially unchanged; or (3) Permanently lost as media obsolesce. (*Choose only one*)

8 Acknowledgements

The research described here has been supported by The Andrew W. Mellon Foundation and Columbia University. Insofar as foundations and universities can be said to have views, the views expressed herein are not necessarily those of the Foundation or the University. Each view expressed is or, at least, has been held by at least one among the authors. The first author acknowledges support from Columbia University, through a contract with Tantalus Inc., from SCILS, Rutgers University, and from the Fulbright Foundation, and Ragnar Nordlie and the Journalism, Library and Information Science Department of the Oslo University College, Norway, for hospitality during the period when this paper was reduced to final form. At Columbia the authors are indebted to many individuals in the Libraries, in Academic Information Systems, and in the academic departments for their participation, encouragement, and cooperation throughout the project period. Elaine Sloan, University Librarian, was critical to the formulation of the project at the outset and an insightful supporter during the project's term. Walter Bourne, David Millman and Gordon Dahlquist were particularly important to the process of creating the online books and various online questionnaires. Lynn Jacobsen Rohrs was a key project participant as the analyst of the web server data. Kate Wittenberg of Columbia University Press, Leo Balk of Garland Press, and Ursula Bollini of Oxford University Press were responsible for providing us with books from their presses and for sharing their insights into the publishing business and critical issues for our research.

9 Biographical Notes.

Paul Kantor is Professor in the Department of Library and Information Science of the School of Communication, Information and Library Studies at Rutgers University, where

he also directs the Rutgers Distributed Laboratory for Digital Libraries, and the Alexandria Project Laboratory for study of the library function. Mary Summerfield is an independent consultant in the information industry based in Oak Park, IL. During the period of this study, she was a Project Director at Columbia University with primary responsibility for the Online Books Evaluation Study. Carol A. Mandel is Dean of Libraries at New York University and Publisher, New York University Press. During the period of this study, she was Deputy University Librarian at Columbia University.

10 References And Literature cited

- Malcolm Getz, *Electronic Publishing in Academia: An Economic Perspective*, **The Serials Librarian** (The Haworth Press, Inc.), Vol. 36, No. 1-2, 1999, pp.263-300.
- Carol Mandel and Mary Summerfield, *Scholarly Monographs Online: Potentialities and Realities Suggested by the Columbia University Online Books Evaluation Project*, January 1998. Expansion of talk given at AAUP and ARL-sponsored conference, *The Specialized Scholarly Monograph in Crisis or How Can I Get Tenure if You Won't Publish My Book? Economics of the Specialized Monograph*, September 11, 1997. <http://www.arl.org/scomm/epub/papers/mandel.html>
- Barbara Kline Pope, *National Academy Press: A Case Study*, **The Journal of Electronic Publishing**, Vol. 4, No. 4, June 1999.
- Robert Nozick, **The Nature of Rationality**. (Princeton University Press) Princeton, 1993. 226pp.
- Carl Shapiro, Hal R. Varian. **Information rules : a strategic guide to the network economy**. (Harvard Business School Press) 1999. Boston, Mass. 352 pp
- James Shapiro, *Saving "Tenure Books" From a Painful Demise*, **The Chronicle of Higher Education**, November 1, 1996, page B6
- Mary Summerfield and Paul B. Kantor, *Columbia's Online Books Evaluation Project: Analytical Principles and Design*, May 1996 revision, <http://www.columbia.edu/cu/libraries/digital/olbdocs/protocol/>
- Mary Summerfield, *Online Books: What Role Will They Fill for Users of the Academic Library?*" in **Finding Common Ground**. Cheryl LaGuardia and Barbara A. Mitchell, editors. New York: Neal-Schuman Publishers, 1998, 313-325. (Paper prepared winter 1996 and delivered at Harvard College in March 1996.) <http://www.columbia.edu/cu/libraries/digital/texts/paper/>
- Mary Summerfield, *Issues in the Economics of Scholarly Communication. A White Paper Supporting The Andrew W. Mellon Foundation-Funded Projects – The Online Books Evaluation Project & Columbia International Affairs Online*, revised March 1998.
- Sanford G. Thatcher, *The Crisis in Scholarly Communication*, **The Chronicle of Higher Education**, March 3, 1995, page B1.
- Marlie Wasserman, *How Much Does It Cost to Publish a Monograph and Why?* Presented at AAUP and ARL-sponsored conference, *The Specialized Scholarly Monograph in Crisis or How Can I Get Tenure if You Won't Publish My Book? Economics of the Specialized Monograph*, September 11, 1997.

ⁱ We also studied the economics of scholarly communication in general, the economics of scholarly publishing, and potential new market models for scholarly books with online books one of the offered formats.

ⁱⁱ Roy Heinz, University of Pennsylvania, pc.