Lynn Silipigni Connaway
Consulting Research Scientist

Clifton Snyder
Software Engineer

OCLC Research

Transaction Log Analyses of Electronic Book (E-book) Usage

Note: This is a pre-print version of a paper published in *Against the Grain* (February 2005, 85-89). Please cite the published version; a suggested citation appears below.

Suggested citation:
Connaway, Lynn Silipigni, and Clifton Snyder. 2005. "Transaction Log Analysis of Electronic (E-Book) Usage." *Against the Grain*, February:85-89.

INTRODUCTION

There has been much discussion of electronic books (e-books) in the literature since the late 1990s. Some believe e-books will replace print books,[1] while others tout that e-books are dead and will never become a format of choice. Stephen Abram believes that the death of the e-book has been exaggerated and that e-book use is dictated by the situation and purpose for seeking the information and by the type of content accessed.[2] Walt Crawford expresses similar beliefs about the contradictions of death vs. overwhelming phenomenon associated with the adoption and viability of e-books.[3] Martha Whittaker and Daniel P. Halloran state that print will always have a place, but that there is also a place for electronic materials, especially reference and scholarly materials.[4] People continue to make predictions about the e-book. Ronaleen Roha and Courtney McGrath report, "...only 43,000 of the most curious among us have bought the [e-book] devices, though researchers predict anywhere from 2.6 million to 28 million in use by 2005."[5]

In order to determine if and how e-books are used, librarians and publishers have queried both users and consumers of e-books. There are several studies that compare the usage of a title that is available to users in both print and e-book format.[6] Other studies analyze e-book usage reports and user surveys to better understand what e-books are being accessed and how users perceive their e-book experience.[7]

Another possible methodology for identifying the e-books that users are accessing and how they are accessing them is transaction log analysis, which can be integrated with other data collection methodologies, including those mentioned above. This type of analysis allows the researcher to unobtrusively identify user search and retrieval patterns and to evaluate systems. Transaction log analysis provides both macro analysis, an analysis of aggregate use data and patterns, and microanalysis, an analysis of individual search patterns. The data can be used to develop systems and services based on user behavior.

There are limitations to transaction log analysis. The users may not be identifiable; therefore, it is usually impossible to associate user demographics with usage patterns or to determine where they access the resource or how and why they use the resource. The transaction logs provide massive amounts of data to manipulate. The types of possible analyses are dependent upon the data collected and stored in the system and when systems change, the data collected may change or no longer be available for analysis. The data collected varies with each system, making comparative analyses between different systems difficult or impossible. There is also some question of the invasion of privacy since the users are not informed that they are being observed[8].

E-book transaction log analysis can provide data about the items being accessed and the types of searches being conducted by the users. The number of accesses of each item and the number of times each screen or page is accessed can be collected and analyzed. This type of analysis can also identify patterns of access, such as the country, date, and time that the user accessed the e-book and the length of time the user spends in the e-book and on the site. Examination of transaction logs provides data about the movements of the user within the site and the e-books, which can assist with system and interface design. Although transaction log analysis can provide data to better determine what e-books are being accessed and how they are being accessed, there is little included in the literature utilizing this methodology for e-books. Don Litzer and Andy Barnett examined Web logs to determine usage of local history e-books in a public library, however, the data reflect usage patterns for a very specific local collection.[9]

SCOPE

The data analyzed for this study were collected from the netLibrary site in 2002, 2003, and 2004. netLibrary is a division of OCLC Online Computer Library Center, Inc. and provides full-text e-books in broad subject areas, including a reference collection. The data and findings reported are a macro analysis. Since the transaction logs provide massive amounts of information, a sample of data was collected for February 26 of 2002, 2003, and 2004, in addition to the annual data. Not all of the data were available for each year since some of the data was not collected initially and other data were no longer relevant after site and system changes.

METHODOLOGY

The intent of the research was to identify the:

- number of users accessing the site,
- number of e-book sessions,
- time of day when users were accessing e-books,
- length of time they were spending on the site and within an e-book,
- number of e-books viewed per session,
- number of pages viewed per session,

- types of searches and frequency,
- search terms used and frequency, and
- users' library type affiliation.

A user is one individual, documented by a specific account ID, but who is not identifiable by any demographic factors. A unique user means that a user is only calculated one time per activity or session per day, based on the account ID. Users can be associated with a type of library, which are defined as academic, public, or "other", where "other" includes government, corporate, special, and K-12 libraries. A session is determined by calculating the time a user accesses the site or the time of manual login through the time of manual or automatic logoff.

RESULTS

Number of Users and Sessions

The number of unique users on February 26 for the three consecutive years increased each year. There were 3,796 unique users on this day in 2002, who viewed 3,268 books (average of 1.3). Eight thousand seven hundred eighty-nine users spent an average of 10.9 minutes per session, with a median time of 3.1 minutes per session, and viewed 7,543 books (average of 1.4) on this day in 2003. On February 26, 2004, 14,350 users spent an average of 10.8 minutes per session, with a median time of 3.5 minutes per session, and viewed 12,291 books (average of 1.5). The average and median minutes per session did not vary much on February 26, 2003 and 2004. There were 10,874 sessions on February 26, 2003 and 15,478 on February 26, 2004. Session data were not available for February 2002.

Unique Pages

The number of unique pages viewed per book and unique pages viewed per session for February 26, 2002, 2003, and 2004 was calculated. A unique page refers to the first time a page of an e-book was viewed. If a page was viewed multiple times by the same account ID, it was only counted one time. On this day, the average number of unique pages viewed in 2002 was 14.1 and the median was 6 pages; in 2003, the average was 16.4 and the median was 7 pages; and in 2004, the average was 18.1 and the median was 6 pages. The average number of unique pages viewed per session for February 26, 2003 and 2004 was 12.8 and 15.5 respectively. The median number of unique pages viewed per session for both of these days was 6. The session data were not available for February 26, 2002. There was a slight increase in the average number of unique pages viewed per book and per session, however, the median was consistent for this date.

Time of Access

The times the site was accessed from all time zones were documented and calculated to Mountain Standard Time (MST), using time zone information from www.worldtimezone.com. The peak usage times were between 10 am and 3 pm MST. See Figure 1.

Search Terms

The majority of searches are keyword searches and most of the search terms are single words. Many of the terms relate to computer science and programming languages or to publishers who provide these materials. Genre searches and prolific authors, such as Shakespeare and King, which most likely refers to Stephen King, are also represented in the top ten search terms. See Tables 1, 2, and 3.

E-Book Subjects Accessed

The subjects of the books accessed are very similar between the different types of libraries. This does not, however, mean that the titles accessed within these subject areas are the same titles. The titles vary by type of library user. Social sciences, science, and technology were consistently accessed on February 26 in 2003 and 2004, as well as for the aggregate for 2003 and 2004. See Tables 4, 5, 6 and 7. These data were not available for 2002. When calculating the top twenty-five books viewed by type of library, 92% were accessed by users affiliated with academic libraries and 8% were accessed by users affiliated with public libraries. There are no users affiliated with other types of libraries who viewed titles that were within the top twenty-five. These data correspond with the percentage of academic and public libraries that are members of netLibrary.

DATA INTERPRETATION

From 2002 through August 2004, there was an increase in the number of netLibrary users and the number of netLibrary sessions. There was little difference in the duration of sessions, books viewed per account, pages viewed per book, or pages viewed per session. In fact, the average time per session is approximately eleven minutes, which is a little less than the average fifteen minutes per book reported in the Columbia University Online Books Evaluation Project.[10] This indicates that the netLibrary e-book collection is being used as a reference collection and the e-books are not being read from cover-to-cover. Since netLibrary did not provide the capability to download the e-books to PDAs or other devices during the study period, this seems to be logical.

The peak usage times for the e-book collection correlate with the time most libraries are open and available, which suggests that convenience may be more of a factor in e-book use than the availability of the library or the time of day. The majority of searches are keyword searches and the subjects of the e-books viewed by users of all types of libraries are similar. Although the subjects of the titles viewed are similar, the titles viewed are not necessarily the same titles within these subject areas. The highest accessed titles are science, social sciences, and technology titles, which correlates with Gordon Coleman's netLibrary usage statistics. He reported, "Of the top 50 titles [accessed], 40 are techie or computer books."[11]

LIMITATIONS OF THE STUDY

The netLibrary transaction logs were not intended for identifying and analyzing the behaviors of the users, but to provide information about the load on the system and to troubleshoot when problems arise. Although the transaction logs provide a wealth of data, a researcher must sift through the data to determine what is useful for identifying user behaviors. This process is very time consuming.

As mentioned above, some data were not collected initially. This is the case for unique session identifiers, which were not logged prior to 2003. This required using the account ID to determine when a session began and ended. Using the account ID to determine a session can be misleading. If a user logs in at a public workstation, but fails to logoff of the workstation and another user accesses netLibrary at the workstation, with the previous user's account ID, the transaction log will associate the entire session to one user, the first one.

netLibrary timestamps every record that is logged in the database, but all accesses are logged in Mountain Time since that is the time zone for netLibrary's location. This becomes a problem when trying to determine relative global access times. For example, if a user in Athens, Greece logged on to the site at 2:35 PM Eastern European Time (EET), the system would record the time as 5:35AM MST. This requires converting the system time to the user's local time.

The data are not formatted before they are loaded into the database, which means that the data need to be normalized This becomes very time consuming when dealing with massive data files.

CONCLUSION

Although there are limitations to the transaction log analysis methodology, it is an unobtrusive method of identifying the behaviors of e-book users. The data can be used to make collection decisions, such as what types of e-books to acquire and what types of print books to digitize. Tracking and interpreting use patterns can provide information for the development of more user-focused systems. It would be beneficial to continue to collect and analyze transaction log data from the netLibrary system to identify what e-books are accessed and when they are accessed. These data could provide a historical analysis of the evolution of e-book adoption and usage.

Figure 1

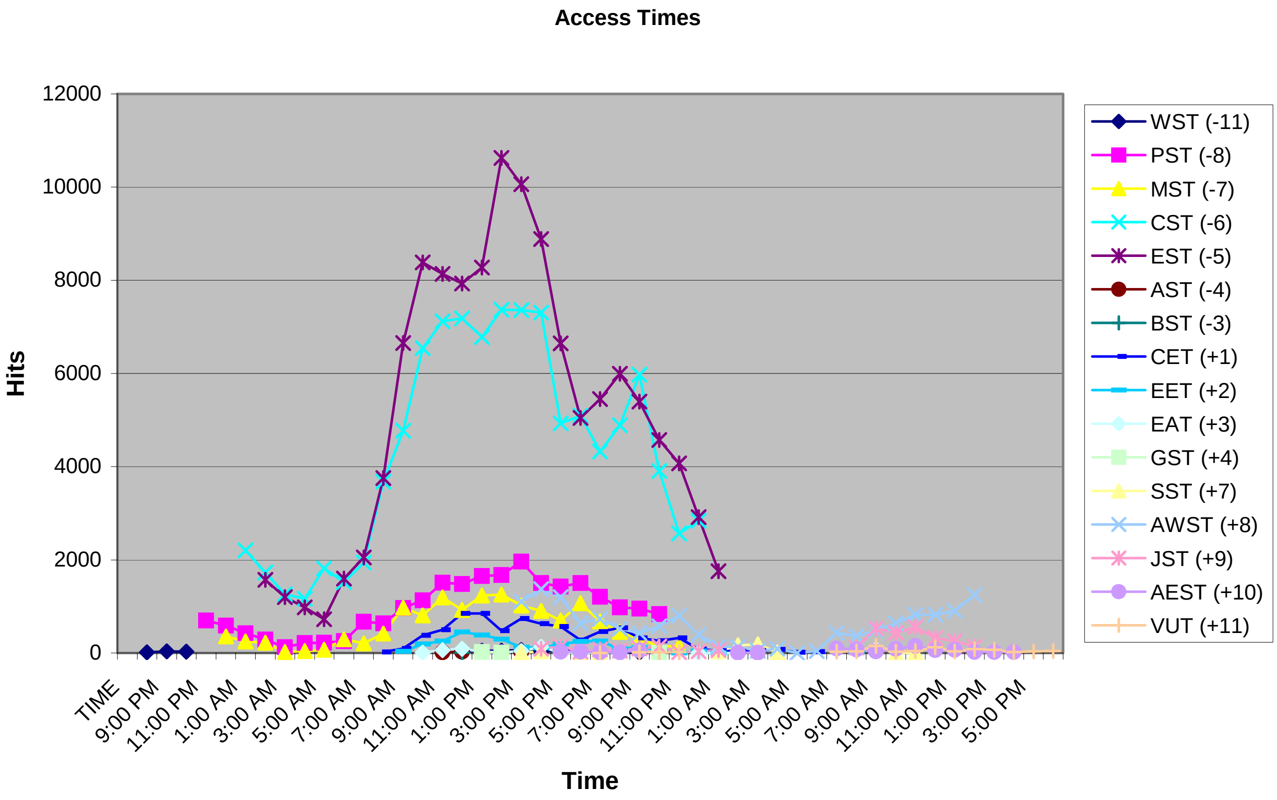
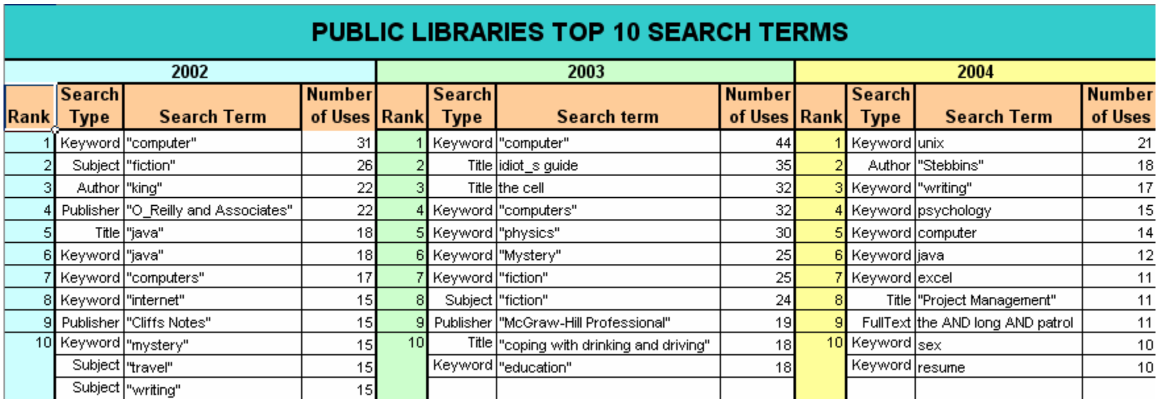


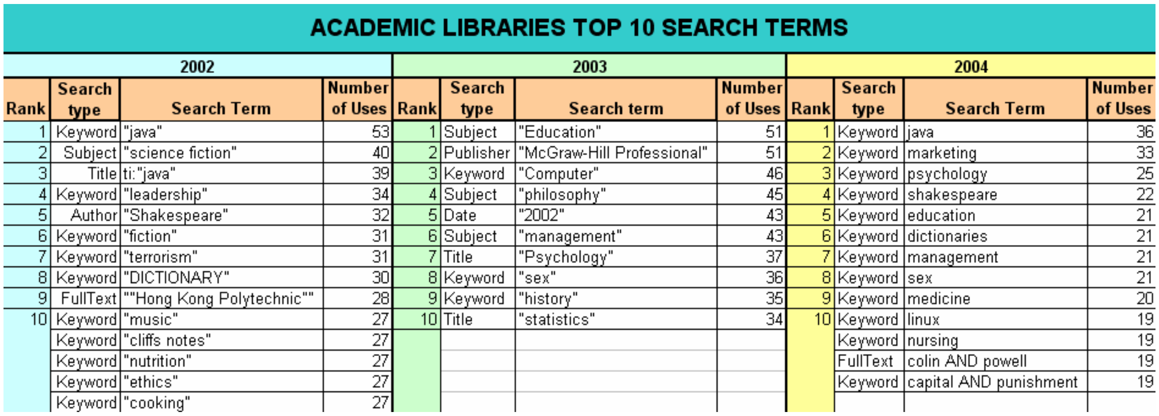
Table 1



PUBLIC LIBRARIES TOP 10 SEARCH TERMS

2002				2003				2004			
Rank	Search Type	Search Term	Number of Uses	Rank	Search Type	Search term	Number of Uses	Rank	Search Type	Search Term	Number of Uses
1	Keyword	"computer"	31	1	Keyword	"computer"	44	1	Keyword	unix	21
2	Subject	"fiction"	26	2	Title	idiot_s guide	35	2	Author	"Stebbins"	18
3	Author	"king"	22	3	Title	the cell	32	3	Keyword	"writing"	17
4	Publisher	"O_Reilly and Associates"	22	4	Keyword	"computers"	32	4	Keyword	psychology	15
5	Title	"java"	18	5	Keyword	"physics"	30	5	Keyword	computer	14
6	Keyword	"java"	18	6	Keyword	"Mystery"	25	6	Keyword	java	12
7	Keyword	"computers"	17	7	Keyword	"fiction"	25	7	Keyword	excel	11
8	Keyword	"internet"	15	8	Subject	"fiction"	24	8	Title	"Project Management"	11
9	Publisher	"Cliffs Notes"	15	9	Publisher	"McGraw-Hill Professional"	19	9	FullText	the AND long AND patrol	11
10	Keyword	"mystery"	15	10	Title	"coping with drinking and driving"	18	10	Keyword	sex	10
	Subject	"travel"	15		Keyword	"education"	18		Keyword	resume	10
	Subject	"writing"	15								

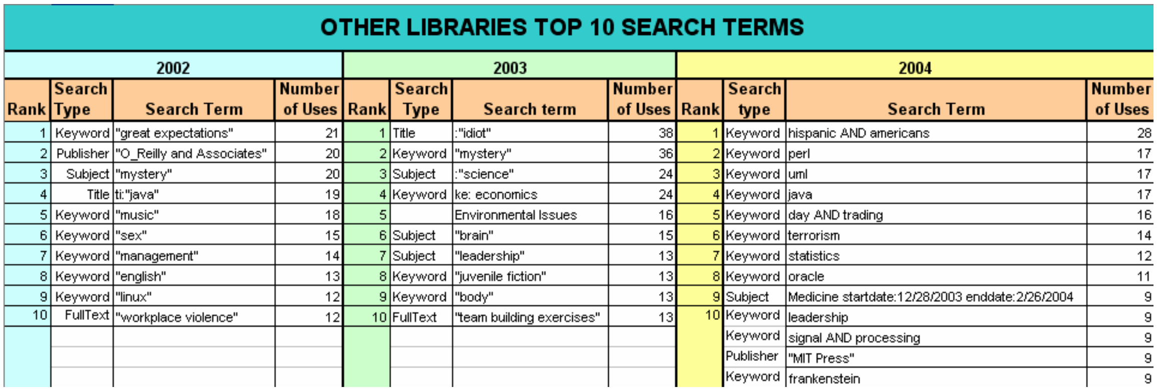
Table 2



ACADEMIC LIBRARIES TOP 10 SEARCH TERMS

2002				2003				2004			
Rank	Search type	Search Term	Number of Uses	Rank	Search type	Search term	Number of Uses	Rank	Search type	Search Term	Number of Uses
1	Keyword	"java"	53	1	Subject	"Education"	51	1	Keyword	java	36
2	Subject	"science fiction"	40	2	Publisher	"McGraw-Hill Professional"	51	2	Keyword	marketing	33
3	Title	ti:"java"	39	3	Keyword	"Computer"	46	3	Keyword	psychology	25
4	Keyword	"leadership"	34	4	Subject	"philosophy"	45	4	Keyword	shakespeare	22
5	Author	"Shakespeare"	32	5	Date	"2002"	43	5	Keyword	education	21
6	Keyword	"fiction"	31	6	Subject	"management"	43	6	Keyword	dictionaries	21
7	Keyword	"terrorism"	31	7	Title	"Psychology"	37	7	Keyword	management	21
8	Keyword	"DICTIONARY"	30	8	Keyword	"sex"	36	8	Keyword	sex	21
9	FullText	""Hong Kong Polytechnic""	28	9	Keyword	"history"	35	9	Keyword	medicine	20
10	Keyword	"music"	27	10	Title	"statistics"	34	10	Keyword	linux	19
	Keyword	"cliffs notes"	27						Keyword	nursing	19
	Keyword	"nutrition"	27						FullText	colin AND powell	19
	Keyword	"ethics"	27						Keyword	capital AND punishment	19
	Keyword	"cooking"	27								

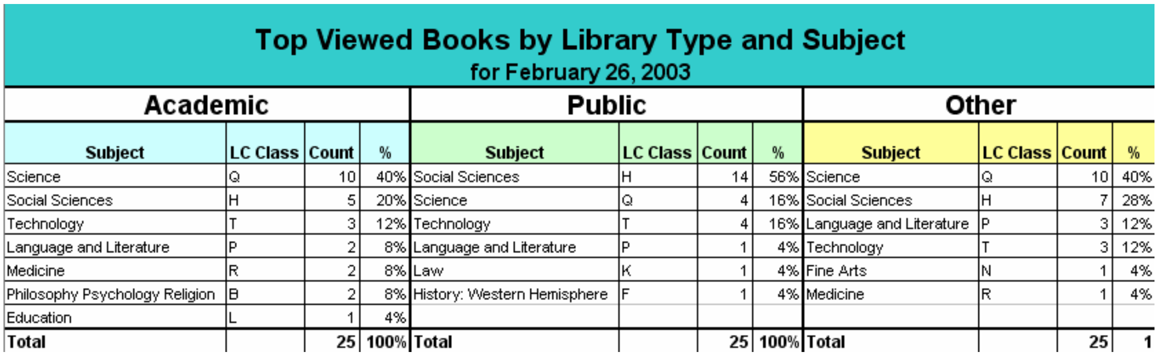
Table 3



OTHER LIBRARIES TOP 10 SEARCH TERMS

2002				2003				2004			
Rank	Search Type	Search Term	Number of Uses	Rank	Search Type	Search term	Number of Uses	Rank	Search type	Search Term	Number of Uses
1	Keyword	"great expectations"	21	1	Title	:"idiot"	38	1	Keyword	hispanic AND americans	28
2	Publisher	"O_Reilly and Associates"	20	2	Keyword	"mystery"	36	2	Keyword	perl	17
3	Subject	"mystery"	20	3	Subject	:"science"	24	3	Keyword	uml	17
4	Title	ti:"java"	19	4	Keyword	ke: economics	24	4	Keyword	java	17
5	Keyword	"music"	18	5		Environmental Issues	16	5	Keyword	day AND trading	16
6	Keyword	"sex"	15	6	Subject	"brain"	15	6	Keyword	terrorism	14
7	Keyword	"management"	14	7	Subject	"leadership"	13	7	Keyword	statistics	12
8	Keyword	"english"	13	8	Keyword	"juvenile fiction"	13	8	Keyword	oracle	11
9	Keyword	"linux"	12	9	Keyword	"body"	13	9	Subject	Medicine startdate:12/28/2003 enddate:2/26/2004	9
10	FullText	"workplace violence"	12	10	FullText	"team building exercises"	13	10	Keyword	leadership	9
									Keyword	signal AND processing	9
									Publisher	"MIT Press"	9
									Keyword	frankenstein	9

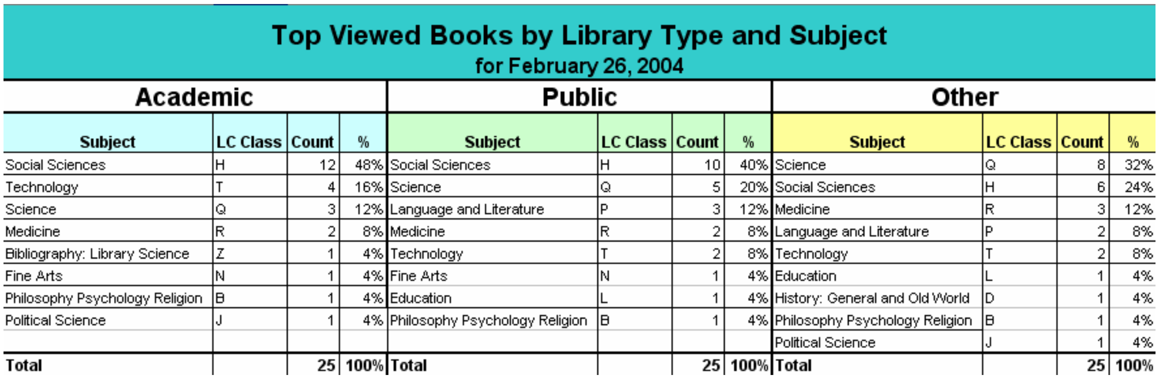
Table 4



Top Viewed Books by Library Type and Subject
for February 26, 2003

Academic				Public				Other			
Subject	LC Class	Count	%	Subject	LC Class	Count	%	Subject	LC Class	Count	%
Science	Q	10	40%	Social Sciences	H	14	56%	Science	Q	10	40%
Social Sciences	H	5	20%	Science	Q	4	16%	Social Sciences	H	7	28%
Technology	T	3	12%	Technology	T	4	16%	Language and Literature	P	3	12%
Language and Literature	P	2	8%	Language and Literature	P	1	4%	Technology	T	3	12%
Medicine	R	2	8%	Law	K	1	4%	Fine Arts	N	1	4%
Philosophy Psychology Religion	B	2	8%	History: Western Hemisphere	F	1	4%	Medicine	R	1	4%
Education	L	1	4%								
Total		**25**	**100%**	**Total**		**25**	**100%**	**Total**		**25**	**1**

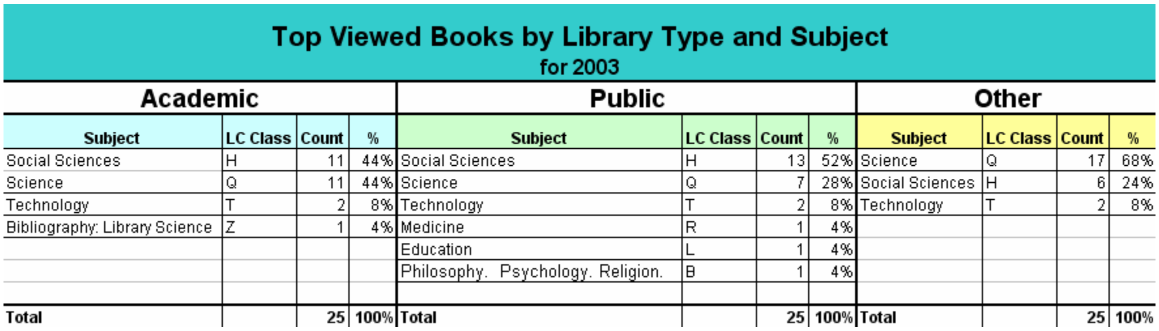
Table 5



Top Viewed Books by Library Type and Subject
for February 26, 2004

Academic				Public				Other			
Subject	LC Class	Count	%	Subject	LC Class	Count	%	Subject	LC Class	Count	%
Social Sciences	H	12	48%	Social Sciences	H	10	40%	Science	Q	8	32%
Technology	T	4	16%	Science	Q	5	20%	Social Sciences	H	6	24%
Science	Q	3	12%	Language and Literature	P	3	12%	Medicine	R	3	12%
Medicine	R	2	8%	Medicine	R	2	8%	Language and Literature	P	2	8%
Bibliography: Library Science	Z	1	4%	Technology	T	2	8%	Technology	T	2	8%
Fine Arts	N	1	4%	Fine Arts	N	1	4%	Education	L	1	4%
Philosophy Psychology Religion	B	1	4%	Education	L	1	4%	History: General and Old World	D	1	4%
Political Science	J	1	4%	Philosophy Psychology Religion	B	1	4%	Philosophy Psychology Religion	B	1	4%
								Political Science	J	1	4%
Total		**25**	**100%**	**Total**		**25**	**100%**	**Total**		**25**	**100%**

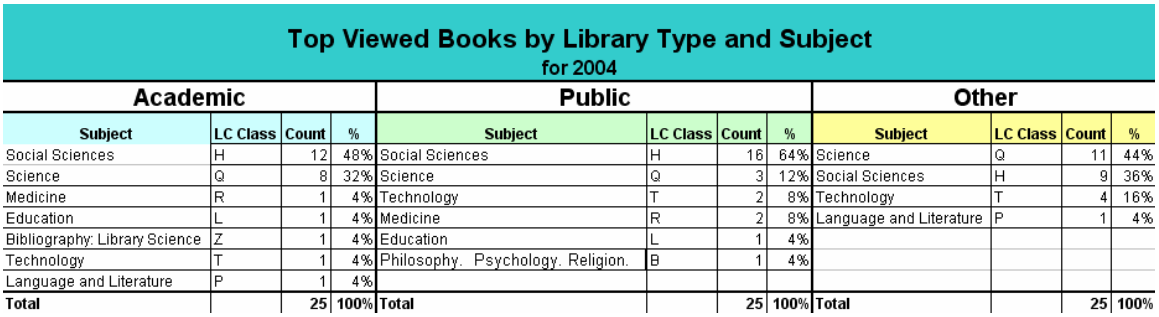
Table 6



Top Viewed Books by Library Type and Subject
for 2003

Academic				Public				Other			
Subject	LC Class	Count	%	Subject	LC Class	Count	%	Subject	LC Class	Count	%
Social Sciences	H	11	44%	Social Sciences	H	13	52%	Science	Q	17	68%
Science	Q	11	44%	Science	Q	7	28%	Social Sciences	H	6	24%
Technology	T	2	8%	Technology	T	2	8%	Technology	T	2	8%
Bibliography: Library Science	Z	1	4%	Medicine	R	1	4%				
				Education	L	1	4%				
				Philosophy. Psychology. Religion.	B	1	4%				
Total		**25**	**100%**	**Total**		**25**	**100%**	**Total**		**25**	**100%**

Table 7



Top Viewed Books by Library Type and Subject
for 2004

Academic				Public				Other			
Subject	LC Class	Count	%	Subject	LC Class	Count	%	Subject	LC Class	Count	%
Social Sciences	H	12	48%	Social Sciences	H	16	64%	Science	Q	11	44%
Science	Q	8	32%	Science	Q	3	12%	Social Sciences	H	9	36%
Medicine	R	1	4%	Technology	T	2	8%	Technology	T	4	16%
Education	L	1	4%	Medicine	R	2	8%	Language and Literature	P	1	4%
Bibliography: Library Science	Z	1	4%	Education	L	1	4%				
Technology	T	1	4%	Philosophy. Psychology. Religion.	B	1	4%				
Language and Literature	P	1	4%								
Total		**25**	**100%**	**Total**		**25**	**100%**	**Total**		**25**	**100%**

[1] Erin Burt, “Farewell to Books,” *Kiplinger's Personal Finance* 55, no.8 (2001): 20.

[2] Stephen Abram, “E-books: Rumors of Our Death Are Greatly Exaggerated,” *Information Outlook* 8, no.2 (2004): 14-5.

[3] Walt Crawford**, “**The White Queen Strikes Again: An Ebook Update,”
Econtent 25, no.11 (2002): 46-7.

[4] Martha Whittaker and Daniel P. Halloran, “The Future of the Book,” Against the Grain 16, no.5 (2004): 38-9.

[5] Ronaleen R. Roha and Courtney McGrath, “Better Than Books?”, *Kiplinger's Personal Finance* 55, no.7 (2001): 110.

[6] Justin Littman and Lynn Silipigni Connaway, “A Circulation Analysis of Print Books and E-Books in an Academic Research Library,” Library Resources & Technical Services 48, no.4 (2004): 256-62; Lynn Silipigni Connaway, “The Integration and Use of Electronic Books (E-books) in the Digital Library,” in *Computers in Libraries 2002: Proceedings* (Medford, NJ: Information Today, 2002), 18-25; Mary Summerfield, Carol Mandel, and Paul Kantor, “The Online Books Evaluation Project: Final Report” (1999). Accessed December 4, 2004, www.columbia.edu/cu/libraries/digital/olbdocs/finalreport.pdf; Susan Gibbons, “NetLibrary eBook Usage at the University of Rochester Libraries. Version 2” (2001). Accessed December 4, 2004, www.lib.rochester.edu/main/ebooks/analysis.pdf; Mary Summerfield, Carol Mandel, and Paul Kantor, “The Potential of Online Books in the Scholarly World: From the Columbia University Online Books Evaluation Project” (1999): 16. Accessed December 4, 2004, www.columbia.edu/cu/libraries/digital/olbdocs/potential.pdf.

[7] Marc Langston, “The California State University E-book Pilot Project: Implications for Cooperative Collection Development,” *Library Collections, Acquisitions, & Technical Services* 27 (2003): 19-32; Nancy J. Gibbs, “E-books Two Years Later: The North Carolina State University Perspective,” *Against the Grain* 13 (Dec. 2001-Jan. 2002): 22+; California State University Libraries Electronic Access to Information Resources Committee and e-Book Coordinating Team, “E-Book Pilot Project Final Report.” Accessed December 4, 2004, http://seir.calstate.edu/ebook/about/report/section_8-2.shtml; Dennis Dillon, “E-books: The University of Texas Experience, Part 2,” *Library Hi Tech* 19, no. 4 (2001): 350-362; Susan Gibbons, “Growing Competition for Libraries,” *Library Hi Tech* 19, no. 4 (2001): 363-367; Carol Ann Hughes and Nancy L. Buchanan, “Use of Electronic Monographs in the Humanities and Social Sciences,” *Library Hi Tech* 19, no. 4 (2001): 368-37; Gibbons, “NetLibrary eBook Usage.”

[8] For more information about transaction log analysis, see Ronald R. Powell and Lynn Silipigni Connaway, *Basic Research Methods for Librarians*, 4th ed. Westport, CT: Libraries Unlimited, 2004, 66-7.

[9] Don Litzer and Andy Barnett, “Local History in E-Books and on the Web: One Library’s Experience as Example and Model,” *Reference & User Services Quarterly* 43, no. 3 (2004): 248-57.

[10] Summerfield, Mandel, and Kantor, “The Online Books Evaluation Project” www.columbia.edu/cu/libraries/digital/olbdocs/finalreport.pdf; Summerfield, Mandel, and Kantor, “The Potential of Online Books in the Scholarly World” www.columbia.edu/cu/libraries/digital/olbdocs/potential.pdf.

[11] Gordon Coleman, “E-books and Academics: An Ongoing Experiment,” Canadian Library Association 4 (2004): 125.