THE IMPACT OF COLLECTION WEEDING ON THE ACCURACY OF WORLDCAT HOLDINGS

A Master's Research Paper submitted to the
Kent State University School of Library
and Information Science
in partial fulfillment of the requirements
for the degree of Library and Information Science

by

Jeffrey A. Young

July, 2002

Author: Jeffrey A. Young
Title: The Impact of Collection Weeding on the Accuracy of WorldCat Holdings
Semester: Summer
Year: 2002
Advisor: Dr. Marcia L. Zeng

ABSTRACT

OCLC's WorldCat database contains 849 million holdings listings for the purpose of associating WorldCat's 48 million bibliographic records with the 41,000 participating libraries in 82 countries that possess those items. In 2001, OCLC celebrated the 30th anniversary of WorldCat. If libraries haven't been diligent about removing holdings of weeded and lost materials, 30 years is a long time for obsolete holdings to accumulate. As OCLC develops new plans to extend the resource sharing capabilities of WorldCat, the reliability of these holdings becomes increasingly important. To the extent that problems exist, these findings can be used to encourage libraries to be more diligent in removing obsolete holdings or perhaps to justify efforts to develop solutions to keep them current. Because this was a pilot study, the sample was limited to books held by members of the OhioLINK consortium. This allowed the author to compare OCLC's holdings against the consolidated catalog for the consortium. While OhioLINK institutions may not perfectly reflect OCLC's current library membership, most of them can claim a long history with OCLC dating back to its origin as the Ohio College Library Center in 1967. This study finds that overall, 7.69% of OCLC's holdings are obsolete compared to the OhioLINK catalog. Little difference was found between ARL and non-ARL institutions. Non-fiction materials made up the bulk of materials sampled and the error rate of 7% for them was in line with the overall rate. The few fiction items in the sample, however, did show an elevated error rate of 20%. Likewise, sampled materials published prior to 1900 were small in number, but exhibited a high error rate of 27%.

Table Of Contents

List of Tables

Chapter 1

PROBLEM STATEMENT

WorldCat (the OCLC Online Union Catalog) contains over 48 million bibliographic records and 849 million location listings (holdings). These holdings serve to associate WorldCat's bibliographic records with the 41,000 participating libraries in 82 countries that hold those items.[1]

Much effort is put into the quality of bibliographic records in WorldCat, but the quality of holdings data is largely ignored. As OCLC implements plans to increase the functionality and flexibility of WorldCat for resource sharing, the quality of this holdings data becomes increasingly important. OCLC participants are committed to associating their individual library symbol (set their holding) for every WorldCat record in their collection. Likewise, they are expected to remove their holding when items are weeded or lost. But while holdings are typically set automatically during original or copy cataloging, they are unset only when libraries make a special effort.

In 2001, OCLC celebrated the 30th anniversary of WorldCat. If libraries haven't been diligent about removing holdings when they weed their collections,

---

[1] OCLC, *OCLC system statistics [News]*. (Dublin, Ohio: OCLC, 2002, accessed 18 June 2002); available from http://www.oclc.org/news/product/statistics.shtm; Internet.

30 years is a long time for obsolete holdings to accumulate. This study quantifies the degree to which obsolete holdings have accumulated.

To keep the scope of this project manageable, two limitations were placed on the sample. First, the analysis for this study was limited to book materials. The analysis of serials, for example, was judged to be too complicated for this effort. Second, only libraries in the OhioLINK consortium were included in the sample. A simple random sample from the population of 41,000 participating libraries would most likely produce a unique institution for most items in the sample and require enormous effort to locate and query a different catalog for each, even if we assume all these catalogs were readily accessible via the Internet. In contrast, OhioLINK provides a single public Web-based catalog for its member libraries, most of which also happen to be members of OCLC. While OhioLINK libraries may not be completely representative of OCLC's current membership, most can claim a long history with OCLC, dating back to its origin as the Ohio College Library Center in 1967.

To the extent that the findings here are a concern to OCLC and its member libraries, this study can be used to encourage them to be more diligent about removing holdings, or justify the development of solutions for keeping them current.

Chapter 2

LITERATURE REVIEW

What little has been written on the accuracy of holdings data comes in the context of interlibrary loan failure analysis. In 1987, David Everett found that citation verification surprisingly held little benefit for improving interlibrary loan success rates for serials, but holdings verification against a union catalog greatly increased the fill rate of article requests. Everett also found that holdings verification against a union catalog wasn't used often enough.[2] This was long ago, but the finding underscores the central role that union catalogs fulfill in the interlibrary loan process.

Most directly relevant to this study is the inclusion of a "title not owned" category in Scott Seaman's analysis of fulfillment failures among OCLC ILL requests processed by Ohio State University (OSU) during a seven-month period in 1990.[3] Of 7,301 ILL requests, 301 (4.1%) are classified as "title not owned." Of 7,846 photocopy requests, 261 (3.3%) are classified as "title not owned." Also of interest is the fact that only 50% of the incoming requests were filled by OSU,

---

[2] David Everett, "Verification in Interlibrary Loan: a Key to Success?" *Library Journal* 112 (November 1, 1987): 37-40.

[3] Scott Seaman, "An Examination of Unfilled OCLC Lending and Photocopy Requests," *Information Technology and Libraries* 11 (September 1992): 231.

although the reasons for this high rate include factors beyond the scope of this study.

Mary Jackson reports that a panel of ILL users convened in 1990 identified 13 issues for improving ILL service.[4] Among them is the need to reduce the elapsed time between request placement and material receipt. An important factor in this regard is the reduction of requests to lenders that can't be filled. While Seaman's study indicates that "title not found" is not a significant factor in ILL request failures, it is a category that could be systematically addressed without forcing any changes on the ILL process itself.

In 1998, Kate Nevins discussed the origins of the OCLC ILL system from 1979 and noted that libraries were starting to allow patrons to initiate ILL requests directly from OCLC systems such as FirstSearch.[5] As noted by Jane Smith, however, a possible consequence of direct patron request is that patrons are directly exposed to the sloppiness of the ILL process.[6] As this practice becomes more common, accuracy of holdings becomes increasingly important.

Also in 1998 and in relation to the patron ILL trend, Chandra Prabha and Edward O'Neill studied the characteristics of books ILL requests via

---

[4] Mary E. Jackson, "Library to Library: ILL: Issues and Actions," *Wilson Library Bulletin* 65 (February 1991): 104-5.

[5] Kate Nevins, "An Ongoing Revolution: Resource Sharing and OCLC," *Journal of Library Administration* 25, no. 2-3 (1998): 65-71.

OhioLINK.[7] In particular, they found that recently published books are frequently requested. Half of the books were published in the preceding seven years while only 10% were published before 1960.

The concern about holdings accuracy becomes more apparent as Barbara Quint reports in 2000 on OCLC's new strategy called Extended WorldCat, which is likened to an Amazon.com-like model for interlibrary loan.[8] Much work needs to be done to the ILL process in terms of accuracy and efficiency, however, before patrons will experience Amazon.com levels of satisfaction.

---

[6] Jane Smith, "An Examination of the Consequences of Electronic Innovations," *Journal of Interlibrary Loan, Document Delivery & Information Supply* 8, no. 4 (1998): 77.

[7] Chandra Prabha and Edward O'Neill, "Interlibrary Borrowing Initiated by Patrons: Some Characteristics of Books Requested Via OhioLINK," *Annual Review of OCLC Research* (1998).

[8] Barbara Quint.. "OCLC Sets Its New Strategy," *Information Today* 17 (December 2000): 7-8.

Chapter 3

OBJECTIVES

The gap in research to be addressed by this study is the degree to which institutional holdings in the WorldCat database remain set for items that are no longer available in participants' collections. Because special effort is required by libraries to remove obsolete holdings from WorldCat and because they have had 30 years to accumulate, the hypothesis is that the number of obsolete holdings will be high enough to be of concern to OCLC and its member libraries.

DEFINITION OF TERMS

- **ARL (Association of Research Libraries):** A non-profit membership organization of leading research libraries in North America.

- **Holdings:** The holdings for an individual WorldCat bibliographic record is a list of OCLC library symbols for participants that possess the item in their collections.

- **HTML (HyperText Markup Language):** A text-based document format for storing and transmitting information for visual rendering by a Web browser.

- **ILL:** Interlibrary loan (including photocopy requests).

- **Obsolete holding:** An OCLC library symbol associated with a WorldCat bibliographic record for which a corresponding bibliographic record no longer exists in the institution's OPAC.

- **OCLC library symbol:** A three-character code assigned to libraries that participate in the creation of WorldCat and used to associate the member library with individual bibliographic records.

- **OCLC number:** An accession number for bibliographic records in the WorldCat database.

- **Online public access catalog (OPAC):** A search interface that allows library patrons to search a library's collection.

- **Participant:** A library that has contracted with OCLC to maintain their library symbol to indicate their holdings of WorldCat records in their collections.

- **Resource sharing:** The sharing of items in library collections between libraries, which is facilitated by WorldCat holdings (specifically, ILL).

- **Weeding:** The act of removing items from a library's collection.

- **WorldCat (The OCLC Online Union Catalog):** A database of 48 million bibliographic records created and used by OCLC participating libraries.

- **XML (eXtensible Markup Language):** A text-based document format for storing and transmitting information for automated processing.

LIMITATIONS OF THE STUDY

To simplify the accumulation and analysis of data, only holdings from OhioLINK members are included in the sample. Also because of the complexity related to checking serials holdings, this study will be limited to book format materials.

A key assumption of this study is that the OhioLINK OPAC is an accurate reflection of its members' collections. No effort will be made to verify an item's existence beyond its presence in the OhioLINK system.

A further assumption is that materials held in both systems will share the same OCLC number. This assumption might fail if holdings were inconsistently dispersed across duplicate bibliographic records in either system, or if OhioLINK holdings were set on records that lacked an OCLC number where one was available.

On the positive side, all OhioLINK members share a common OPAC vendor and the circulation systems for each are closely synchronized with the OhioLINK OPAC.

Chapter 4

METHODOLOGY

The population for this study was the set of OhioLINK member holdings for book materials in the WorldCat database. From this population, a simple random sample was studied.

The first task to derive the sample was to correlate the OCLC library symbols used in WorldCat holdings records with OhioLINK's 81 institutions.[9] This was done by manually comparing the list of names on OhioLINK's Web site with names in the OCLC library symbol table. In some cases, OCLC library symbols could not be found for some OhioLINK institutions. In other cases, multiple OCLC library symbols were associated with a single OhioLINK institution. In the end, 108 OCLC library symbols were found for 73 of the 81 OhioLINK institutions. A simple random sample of the entire population of book material holdings for these 108 symbols resulted in 1,210 OCLC number/OCLC library symbol pairs (holdings).

The next step was to obtain a list of OhioLINK holdings for each of the OCLC numbers. OhioLINK's Web interface allows users to search by OCLC

---

[9] OhioLINK, *OhioLINK Member Libraries.* (Accessed 18 June 2002); available from http://www.OhioLINK.edu/members-info/mem-links.php.

number and can produce a Web page listing the OhioLINK institutions that hold the item. To expedite the search process, a macro was written to read OCLC numbers from the sample's input file and interact with OhioLINK's Web server directly. For each OCLC number found, the macro wrote the holdings data HTML page returned by OhioLINK to a file for later evaluation. If the OCLC number was not found in OhioLINK, an empty file was created. The file name created for each sample item was a combination of the OCLC number and the target OCLC library symbol that should be represented within when the HTML file was examined. Since the macro could interact with OhioLINK's OPAC at a much faster rate than a human user, a delay loop was inserted to avoid overwhelming the server. In a further effort to minimize the impact on other users, the entire sample was processed overnight to avoid peak usage times.

Writing the software to extract OhioLINK holdings from the HTML pages proved to be a bit more challenging. If the OhioLINK server had returned results in XML, all the data would have been clearly and easily extracted with a computer program. HTML, however, is designed to be rendered for human visual consumption and doesn't necessarily contain clues to guide automated processes. Fortunately, in this case, the HTML pages did contain enough hidden indications to clearly identify the holding institutions.[10] A program was written to

[10] A significant complication for the program was that holdings for CONSORT and OPAL consortia members were treated differently in the HTML from other individual institutions. Even in this case, though, the information was still adequate to resolve the individual institutions.

parse the HTML page and convert the OhioLINK institutions found there into a string of equivalent OCLC library symbols separated by spaces. Next, the program parsed the target OCLC library symbol that was encoded in the HTML filename and searched for its presence in the generated list. If the symbol was found, the OCLC number, the target OCLC library symbol, and the word "GOOD" were written to a log file. If the symbol was not found, the OCLC number, target OCLC library symbol, and the word "OBSOLETE" were written to the log.

Last, a MARC Communications Format record for each OCLC number in the sample was extracted from WorldCat and written to a file. From these bibliographic records, the DATE1, FICT, and LANG fixed fields were extracted and merged into the log to enable breakdowns of the results according to those variables.

Chapter 5

DATA ANALYSIS

The overall finding of the study is that 93 of the 1,210 WorldCat holdings sampled (7.69%) aren't reflected in the OhioLINK catalog and are thus obsolete. Table 1 shows a breakdown of the results by the FICT bibliographic fixed field. The vast majority of obsolete holdings (1,141 of 1,210) are in the non-fiction category, but the percentages indicate that fiction materials are much less carefully weeded with 20.29% obsolete compared to non-fiction materials with 6.92% obsolete.

Table 1

Count of Holding Status	**Holding Status**			
FICT	**GOOD**	**OBSOLETE**	**Grand Total**	Error Rate
Non-Fiction	1062	79	1141	6.92%
Fiction	55	14	69	20.29%
Grand Total	**1117**	**93**	**1210**	7.69%

Table 2 shows the breakdown by decade of the DATE1 bibliographic fixed field. Although materials published prior to 1900 are a small percentage of the sample (83 of 1,210 or 6.86%), the study finds a fairly high rate of obsolete holdings in the group (25 of 93 or 26.88%). Further study is needed to examine the bibliographic records in the group to determine their characteristics.

Table 2

Count of Holding Status	Holding Status			
Decade	**GOOD**	**OBSOLETE**	**Grand Total**	Error Rate
1540	1		1	
1610	1		1	
1640		3	3	100.00%
1650		1	1	100.00%
1660	1	3	4	75.00%
1670	1		1	
1680	1		1	
1690	1	3	4	75.00%
1700	1		1	
1760	1		1	
1770		1	1	100.00%
1800	3	1	4	25.00%
1810	5	10	15	66.67%
1820	3		3	
1830	4		4	
1840	4	1	5	20.00%
1850	4		4	
1860	2		2	
1870	8	2	10	20.00%
1880	6		6	
1890	11		11	
1900	18	2	20	10.00%
1910	13	1	14	7.14%
1920	22	2	24	8.33%
1930	26	1	27	3.70%
1940	35	1	36	2.78%
1950	70	6	76	7.89%
1960	166	5	171	2.92%
1970	196	16	212	7.55%
1980	211	22	233	9.44%
1990	266	11	277	3.97%
2000	34	1	35	2.86%
#N/A	2		2	
Grand Total	**1117**	**93**	**1210**	7.69%

Table 3 shows a breakdown by the bibliographic LANG fixed field. Most materials sampled are in English (1,079 of 1,210 or 89.17%) and thus reflect the general error rate at 7.88%. The other languages aren't sufficiently represented in the sample to derive conclusions about them individually.

Table 3

Count of Holding Status	Holding Status			
LANG	**GOOD**	**OBSOLETE**	**Grand Total**	Error Rate
Ara	1		1	
Chi	2		2	
Dut	1		1	
Eng	994	85	1079	7.88%
Fre	18	3	21	14.29%
Frm	1		1	
Ger	34	1	35	2.86%
Gre	3		3	
Heb	1		1	
Hun	1		1	
Ind	6		6	
Ita	7	2	9	22.22%
Jpn	3		3	
Lat	3	1	4	25.00%
Per	2		2	
Pol	2		2	
Por	1		1	
Roa		1	1	100.00%
Rum	2		2	
Rus	16		16	
Scr	1		1	
Spa	14		14	
Swe	1		1	
Tur	2		2	
Ukr	1		1	
Grand Total	**1117**	**93**	**1210**	7.69%

The sample was not large enough to justify a breakdown of the results by individual institution, but Table 4 shows a breakdown by institution type. The OhioLINK consortium includes five members of the Association of Research Libraries (ARL) group: Case Western Reserve, Kent State University, Ohio State University, Ohio University, and the University of Cincinnati. Obsolete holdings for ARL institutions stands at 7.51%, which is comparable to the 7.06% observed for the non-ARL institutions. The State Library of Ohio is also a member of OhioLINK for whom the table shows a relatively high rate of obsolete holdings at 7 of 20, or 35%. Further study is needed to determine the nature of the bibliographic records in this category.

Table 4

Count of Type Value	Holding Status			
Type Value	**GOOD**	**OBSOLETE**	**Grand Total**	Error Rate
ARL Libraries	419	34	453	7.51%
State Libraries	13	7	20	35.00%
Non-ARL Libraries	685	52	737	7.06%
Grand Total	**1,117**	**93**	**1,210**	7.69%

Chapter 6

CONCLUSION

This pilot study was designed to get a general sense of the error rate in OCLC's holdings information. OCLC's resource sharing services are largely based on this holdings data and the accuracy of the holdings is a key factor in its effectiveness. A simple random sample of book holdings for OhioLINK institutions provides a convenient set of data for analysis.

This study indicates an overall degree of error in WorldCat holdings of 7.69%. Although fiction items made up a small percentage of the total items, they exhibit a relatively high error rate of 20.3%. Items published prior to 1900 were also a small percentage of the items, but likewise showed a relatively high error rate of 26.9%. Breakdowns by language of publication and ARL vs. non-ARL showed no significant differences. While improved accuracy of holdings will not solve the greater inefficiencies of ILL processing mentioned in the literature, it is a problem that can be addressed systematically and with minimal impact on existing processes.

BIBLIOGRAPHY

Everett, David. "Verification in Interlibrary Loan: a Key to Success?" *Library Journal* 112 (November 1, 1987): 37-40.

Jackson, Mary E. "Library to Library: ILL: Issues and Actions." *Wilson Library Bulletin* 65 (February 1991): 102-5.

Nevins, Kate. "An Ongoing Revolution: Resource Sharing and OCLC." *Journal of Library Administration* 25, no. 2-3 (1998): 65-71.

OCLC. *Annual Report 2000/2001*. Dublin, Ohio: OCLC, 2001.

OCLC. *Principles of Cooperation [OCLC – WorldCat]*. Dublin, Ohio: OCLC, 2002. Available from http://www.oclc.org/worldcat/cooperation/principles.shtm. Accessed 18 March 2002.

Prabha, Chandra and Edward O'Neill. "Interlibrary Borrowing Initiated by Patrons: Some Characteristics of Books Requested Via OhioLINK." *Annual Review of OCLC Research* (1998). Available from http://www.oclc.org/research/publications/arr/1998/prabha_oneill/patron.htm. Accessed 18 March 2002.

Quint, Barbara. "OCLC Sets Its New Strategy." *Information Today* 17 (December 2000): 7-8.

Ranganathan, Shiyali Ramamrita. *The Five Laws of Library Science*. Bombay: Asia Pub. House, 1964.

Seaman, Scott. "An Examination of Unfilled OCLC Lending and Photocopy Requests." *Information Technology and Libraries* 11 (September 1992): 229-35.

Smith, Jane. "An Examination of the Consequences of Electronic Innovations." *Journal of Interlibrary Loan, Document Delivery & Information Supply* 8, no. 4 (1998): 71-78.