ENHANCING SUBJECT ACCESS TO MONOGRAPHS IN ONLINE PUBLIC ACCESS CATALOGS: TABLE OF CONTENTS ADDED TO BIBLIOGRAPHIC RECORDS

Vinh-The Lam, M.L.S.

(A Chapter of ***Saving the Time of the Library User Through Subject Access Innovations: Papers in Honor of Pauline Atherton Cochrane*** / edited by William J. Wheeler; published by University of Illinois at Urbana-Champaign in 2000; p. 162-172)

ABSTRACT

Subject access to monographs through online public access catalogs (OPACs) has always been a major concern for large research and/or academic libraries. Academic library practice of providing subject access to monographs has proved inadequate, especially in case of composite works. Many techniques have been proposed to enhance subject treatment of monographs in OPACs. This paper briefly reviews these efforts in the past and presents the case of adding Table of Contents as one of the most useful and probably also one of the most cost-effective ways of improving subject access to monographs in an academic environment.

Several decades ago, Ranganathan affirmed as the First Rule of Library Science that "Books Are For Use." In order to help users make good use of "books", the library catalog should be able to provide a good representation of the books' contents. This goal set for the catalog has always been elusive. The library world has been struggling to make it a reality. With the first-generation online public access catalog (OPAC) coming on the scene in the late 1970s replacing the traditional card catalog, librarians had great hopes for improved access in general and improved subject access in particular. But before long the library community realized that the OPAC was, in this area of subject access, not much netter than its predecessor, with all the same limitations.

INADEQUACIES OF LCSH

By the 1970s, Library of Congress Subject Headings (LCSH) have been used to provide primary subject access in large research and/or academic libraries for some eighty years. In spite of its firmly established place in these libraries, LCSH had serious weaknesses which were identified by various authors and researchers between the 1950s and the late 1970s. These were grouped together with suggested improvements by Pauline Atherton and Monika Kirtland in their 1981 *"Critical Views of LCSH: A Bibliographic Survey."* [1]

Among the most serious weaknesses were: inconsistent form of headings, lack of clarity in subdivision structure, lack of currency of heading terms, and lack of specificity of heading terms. The impediment in OPAC access and retrieval was created not only by these internal weaknesses of LCSH

but also by subject cataloging practices. The limited number of subject headings assigned to monographs as well as the practice of providing a more general subject heading for a monograph that covers several more specific topics resulted in very inadequate subject access to monographs. Even now, LC still instructs its catalogers in its ***Subject Cataloging Manual: Subject Headings*** as follows: *"Assign to the work being cataloged one or more subject headings that best summarize the overall contents of the work and provide access to its most important topics… Assign headings only to topics that comprise at least 20% of the work… The number of headings that are required varies with the work being cataloged. Sometimes one heading is sufficient. Generally a maximum of six is appropriate… Do not assign more than ten headings to a work."* [2] The reality turned out to be that *"an average of only two headings per book is provided…"* [3] Catalog use studies have shown that subject searches account for more than half (52%) of all catalog use, and the failure rate is this category of search is 50%.[4]

EXPLORING WAYS TO IMPROVE SUBJECT ACCESS IN THE OPAC

Pauline Atherton Cochrane, one of the pioneers in the history of information retrieval and library automation, has called the 1970s ***"The Development Decade"*** and the 1980s ***"a time for reassessment."*** [5] A series of research projects were carried out, mostly during the 1980s, with the objective of improving and assessing the online public access catalog.

The most noted and probably most comprehensive project to look into improving subject access for the OPAC was Pauline Atherton's ***Subject Access Project (SAP)*** for the Council on Library Resources in 1978. [6] In this important research work, Atherton and her colleagues tried to provide increased subject access to an online database called BOOKS. Subject terms, selected according to clearly defined rules, from tables of contents and end-of-book indexes were added to bibliographic records in BOOKS. The average number of selections per book was 32.4. The research findings were a clear indication of better subject access.

	MARC	BOOKS
Number of searches	90	90
Number of Relevant Items Retrieved	56	131
In Social Sciences	31	61
In Humanities	25	70
Number of Known Relevant Items NOT retrieved	117	42
Average Precision Rate	35%	46%
Search Time	.27 Hour	.14 Hour
Social Sciences	.15 Hour	.08 Hour
Humanities	.12 Hour	.06 Hour

The most significant finding was that increased recall did not necessarily mean a decreased precision as many other researchers have feared. Basically, Atherton's SAP technique clearly demonstrated four benefits from such a database: greater access, greater precision, less cost, and the ability to answer questions that are otherwise impossible using regular catalog information.

Many research studies in the 1980s explored ways of enhancing subject access to monographs and almost all pointed to the addition of contents terms as the best way. A study by Karen Markey in 1983 found that one of the most desired new capabilities of OPAC was to search a book's table of contents, summary, or index. In 1987, Karen Calhoun and Karen Markey, comparing the search value of some of the content components, found that notes in MARC tags 505 and 520 (Contents Note and Summary Note) accounted for the largest percentage increase in the number of unique words with an average of 15.5 new subject terms per record. An in 1989, Florence DeHart and Karen Matthews confirmed the benefits of searching chapter titles. [7] In a landmark paper from this period, Carol Mandel, having determined the main value of enriched records to be to provide access to parts of books, suggested eleven alternative ways to provide such information. Basically, she envisaged there ways of carrying out these alternatives: creation of separate databases of monographs content indexing, including monograph content indexing in LC MARC records, and adding monograph content indexing in either utility databases or online catalogs. For each of these three ways, she proposed using controlled and uncontrolled vocabulary. For the uncontrolled vocabulary scenario, SAP technique was proposed as the mechanism for content indexing of a select set of monographs. Although she identified three ways of doing this in her matrix, a closer look revealed that, in fact, there were only two ways: either providing monograph content information in separate databases or right in the bibliographic records. In the conclusion of her paper, Mandel said: *"The possible alternatives suggested in this paper need to be tested and weighed against cost-effectiveness and need. Should we enhance the MARC record to improve subject access? We won't know until we try."* [8] People did try. One of the most important experiments in this area in the 1980s reaffirmed the findings from Atherton's research. This experiment, called ***Enhanced Subject Project (ESP)***, was carried out at the Australian Defence Force Academy (ADFA) library. After having considered several methods, ADFA library staff decided to choose Atherton's SAP technique. Supplementary terms selected from tables of contents of books and, where needed, the index, (on an average of twenty to twenty five per book), were into field 653 in the bibliographic records for these books. The first report, after six months into the experiment, with more than 6,000 books receiving ESP, was published in 1988. *"The Enhanced Subject Project has demonstrated that the use of contents terms is a viable and cost-effective technique for dramatically increasing the number of subject access points to the contents of books without a serious increase in false drops."* [9] Five years later, with ESP now applied to close to 40,000 records, or about 25% of ADFA's collection, another report was made public and the benefits of ESP reconfirmed. Halfway around the world, Atherton's SAP technique was put in practice in Sweden. Irene Wormell reported on the application of SAP technique in indexing Swedish government official reports for the SOU database (SOU stands for Statens Offentliga Utredningar = Government Official Reports) at the Library of the University of Lund. Of particular interest was the use of SAP technique in capturing terms used in tables and graphs included in these reports for indexing purposes. Wormell found that the captions of tables and graphs were *"usually more self-explanatory than the fancy titles or chapter headings of the reports."* [10] Wormell concluded that *"SAP-indexing is a new way to produce subject description for monographs and, at a moderate cost, retrieve those specific parts of publications which are not usually accessible in traditional IR-systems."* [11] In the late 1980s, as many American university libraries started to enrich bibliographic records by adding either abstracts or tables of contents, the OCLC Office of Research, using recall and precision as evaluation measurements, conducted a research project to evaluate retrieval effectiveness of an online test database of about 4,900 records, which contained both abstracts and tables of contents. Ten records were randomly selected from the database to generate twenty queries, which were then translated into Boolean form and used to search the database. The searches were done at four levels: level 1 using an index containing title and subject headings fields, level 2 is level 1 plus tables of contents, level 3 is level 1 plus abstracts, and level 4 is level 1 plus both tables of contents and abstracts. The findings showed that recall improved as additional content information was added to the records but this increase in recall was accompanied by a decrease in precision: [12]

	Average	
	Recall	**Precision**
Level 1	0.17	0.71
Level 2	0.26	0.59
Level 3	0.29	0.60
Level 4	0.34	0.60

Recently, at the University of Minnesota Bio-Medical Library, a project was undertaken to enhance online bibliographic records for the reference collection monographs with meaningful tables of contens and summaries. *"Of the 1,100 enhanced records, 33% were enhanced with tables of contents, 39% with summaries only, and 20% with both a table of contents and summary… Overall, for this project, the benefits of adding the content-bearing data to the records outweighs problems with precision*." [13]

DISCUSSING STANDARDS FOR SUBJECT ACCESS ENHANCEMENTS

In 1991, the Library of Congress issued MARBI Discussion Paper No. 42, in which tables of contents were listed as one of the four additions (the other three are: indexes, abstracts, and book reviews) to be considered for OPAC bibliographic records. MARC field 505 (Formatted Contents Note) was singled out as the location for this addition. *"Contents could be represented by titles and possibly authors of chapters or by chapter number, full title, and page numbering."* [14] After responses and comments to this Discussion Paper were received and analyzed, it became clear that the majority of respondents opted to go with tables of contents. MARBI Discussion Paper No. 46, therefore, limited itself in discussing issues related to the enhancement of bibliographic records with table of contents (TOC) information. It provided a good discussion on how to change field 505 to accommodate TOC information. Basically, it recommended that field 505 become a repeatable field with two subfields, $a for authors of chapter titles and $t for chapter titles. It also raised some concerns about name authority control for these enhancements.[15] One of the disadvantages of the addition of TOC information to the bibliographic records is that it makes the latter too lengthy and too hard to read. Recently, the Library of Congress Cataloging-In-Publication (CIP) Office started providing electronic CIP. Part of this program was the provision of lengthy TOC information in separate files which were linked to the bibliographic records through a URL (Uniform Resource Locator) in field 856. These files are now accessible at the following URL: http://lcweb.loc.gov/catdir/toc/ [16] LC seems to follow, though not completely, Mandel's suggestion of providing monograph content indexing in separate databases.

CURRENT DEVELOPMENTS IN TOC

After the release of these two MARBI discussion papers, the main concern for most libraries was the cost involved in implementing these enhancements. Many libraries will not have the money and/or the staff to do this, considering the financial constraints of the 1990s. It appeared that the solution ought to be sought in finding a way to automatically capture content-bearing information for inclusion into the online catalog. The new scanner technology was tested successfully at the School of Library Science and Documentation, Colegio Maximo de Cartuja, Granada, Spain in the mid-1990s. In this research project, the OCR (Optical Character Recognition) scanner technology was used to extract tables of contents from composite monographs. This information was then automatically converted into SGML (Standard Generalized Markup Language) format by a commercially available application called Rainbow

Maker. Finally, these SGML-tagged information components were processed into MARC format for inclusion into the online catalog, using another application designed by the Electronic Text Center of the University of Virginia called "tei2marc." (tei stands for Text Encoding Initiative). The researchers then created two test databases: each containing two hundred records for items in the fields of librarianship, documentation and applied computer science, the first one directly from the regular online catalog, and the second from the enhanced online catalog. Twenty queries were searched against the two test databases. It was found that *"the system with enriched records is more effective in satisfying user needs than the operating one. For fixed recall points, average precision for the catalog system with enriched records is between 21 percent and 45 percent better than that of the operating catalog system. The best results appear for the midrange recall values."* [17]

Just as the cataloging operation was made more efficient through cooperation and/or centralization almost a century ago, the implementation of enhanced subject access with TOC information in individual libraries may be outsourced to vendors. One vendor is currently providing this service at a reasonable price, 1.00 US dollar per record. The technologies helping make this dream come true are: electronic scanning of TOC data by the vendor, and electronic transfer of data from the vendor to the individual libraries via FTP (File Transfer Protocol). This vendor proposes to use the repeatable local field 970 to store TOC data. Recently, it has started to perform authority control for authors' names provided in these fields. Hence, enhanced subject access to monographs in large academic and/or research libraries is now possible at a reasonable cost. A new concern arises though. Claus Poulsen, in his study on the number of composite works in the library collection and the number of articles or citable works in these books, found that *"The proportion of composite works is between 10% and 20%. The number of articles in the composite works varies from 20 to 30 articles per book.... This implies that the libraries under consideration can add access to between 200% and 600% more works to their catalog without buying one more book, just by adding the tables of contents of their composite works."* [18] Academic and/or research librarians now have to find solutions to the potential challenge of information overload. It is really unfortunate that current developments in this area do not use Atherton's SAP technique. Convenience and low cost have been given preference over a good and proven technique. First of all, SAP technique requires some intelligent human intervention while the current TOC services are totally mechanical. The use of scanning technology precludes the elimination of fancy but mostly irrelevant if not misleading terms used in chapter headings in the TOC of many books. And, of course, the index for TOC terms will be swamped with generic and totally useless terms like "foreword", "preface", "introduction", "afterword", etc. taken automatically from the TOC by scanners. With SAP technique all of these useless terms would be eliminated. Secondly, SAP technique not only makes use of terms from the TOC but also includes terms taken from the index of books, which are far more content-specific. To illustrate these points, let's take a look at how SAP technique was actually used for one book in the famous 1978 project conducted by Atherton:

> Turney-High, Harry Holbert, 1899-
> Primitive war: its practice and concepts / by Harry Holbert Turney-High; foreword by David C. Rapoport; afterword by Harry Holbert Turney-High. -- 2nd ed. -- Columbia, S.C.: University of South Carolina Press, 1971.

Its one-page TOC looks like this:

Contents

Foreword

Preface to the First Edition

Part I. The Practice of Primitive War

CHAPTER PAGE

The index has a total of 11 pages (pp. 278-288).

The following represents the record for this items in BOOKS:

RSN - 00670542
SNO - 0978
CCN - GN497 T87 1971
ME - Turney-High, Harry Holbert, 1899-
TI - Primitive war; its practice and concepts. Foreword by David C. Rapoport. Afterword By Harry Holbert Turney-High.
IM - Columbia, University of South Carolina Press, 1971
COL - 288 p.
PY - 1971
LCH - War
CT - *PRACTICE OF PRIMITIVE WAR (P. 5-140) ; WEAPONS (P. 5-20) ; THEORY OF WAR (P. 21-38) ; FORMATIONS (P. 39-60) ; DISCIPLINE AND COMMAND (P. 61-90) ; FUNCTIONAL DESIDERATA (P. 91-106) ; INTELLIGENCE SURPRISE AND COUNTERSURPRISE (P. 107-122) ; BATTLE PLANS (P. 123-140)
CT - *CONCEPTS OF PRIMITIVE WAR (P. 141-253) ; SOCIO PSYCHOLOGICAL MOTIVES (P. 141-168) ; ECONOMIC MOTIVE (P. 169-186) ; MILITARY VALUES

(P. 187-204) ; MILITARY ATTITUDES (P. 205-226) ; WAR AND ORGANIZATION OF SOCIETY (P. 227-253)
CT - *SURVIVAL AND REVIVAL OF PRIMITIVE WAR (P. 254-265)
IT - AFRICA ; DAHOMEAN ; ENEMY LIFE VALUATION OF (P. 220-226) ; EURASIA
IT - GIBEAH CAMPAIGN AGAINST (P. 32-38) ; GREAT PLAINS AMERICAN
IT - HEAD TAKING (P. 196-200)
IT - HEBREWS EARLY MILITARY OPERATIONS (P. 31-38) ; IROQUOIS ; JIBARO
IT - LAND ECONOMIC (P. 182-186) ; LIFE ATTITUDES REGARDING (P. 210-215)
IT - MELANESIA ; METHODS TACTICAL (P. 21-38) ; OCEANIA ; OJIBWAY
IT - OMAHA ; PLANS PRINCIPLE OF SIMPLICITY OF (P. 123-137) ; POLYNESIANS
IT - SAMOAN ; SEX MOTIVE (P. 151-164) ; SLAVING (P. 178-182)
IT - SOCIAL ORGANIZATION (P. 227-253) ; SOCIOLOGY (P. 227-253)
IT - TENSION RELEASE (P. 141-145) ; WOMEN (P. 151-164) ; ZULU [19]

(CT = Contents Terms from the TOC; IT = Index Terms from the Index)

We can see clearly that useless terms (foreword, preface, afterword, bibliography, index) and non-specific terms, such as "form and function", "military horizon", "certain" from the TOC chapter headings were dropped in the Contents Terms (CT). At the same time, the Index Terms (IT) really provided highly content-specific subject access points. There should be no doubt about the superiority of SAP technique.

SMALL SURVEY AT TWO CANADIAN UNIVERSITIES

The author of this paper did a search on the OPACs of University of Saskatchewan (U of S, without TOC) and University of Western Ontario (U of WO, with TOC) on August 12, 1998. The main reason for choosing these universities is that they are quite similar in terms of monograph holdings (1.5 million and 2.1 million respectively) [20] and holdings per medical student (266 and 267 respectively) [21]. The search was done in the Keyword Index, and the search term was "Prostate Cancer." The search result was: 16 items were retrieved from the U of S OPAC, and 101 from the U of WO OPAC. Only two (2) items appeared on both lists. Of the 99 items on the U of WO list that do not appear on the U of S list, 74 do not contain the term "Prostate" in their titles. A second search on the U of S catalog, using Title Index, for these 74 items revealed that U of S had 26 of them. The 26 items were retrieved in the first search because U of WO catalogue records had TOC information. If the U of S OPAC had TOC information, a keyword search would have retrieved 42 items, an increase of 163%.

CONCLUSION

It is too early for a final assessment of subject access enhancement by TOC in OPACs. Some academic libraries that have TOC in their OPACs, however, report some positive preliminary results. *"Recall has obviously been greatly enhanced ... Overall, I think students are glad that we have TOC information ... They can see what it is in a book and even how many pages each chapter covers. That has been very much appreciated by many students whom I have helped..."* [22] *"...That it is a way of being more precise, and is helpful for those users to work from their office, and identify before going to the library, which library materials are worth examining."* [23]

Some decisions have to be made regarding the indexing of this information. Most libraries have decided to have section/chapter title information put in both keyword and title indexes. With search words found in TOC data now highlighted in most systems, it is no longer confusing for users searching the title index. At the Hong Kong University it was decided to have a separate index for TOC. These libraries also

have witnessed some impact on their cataloging procedures. *"The chief impact for cataloging procedures has been in ensuring that staff are aware of the span of records (based on CAT DATE) that have been extracted and sent for TOC processing. If changes are to be made to records being processed for TOC, staff need to wait or repeat the changes once the records are reloaded into the database as they overlay the existing bibliographic records. We have 'protected' all TOC 9XX fields so that they cannot be erased by any future overlay process"* [24] Subject access enhancement with complete TOC information is now a reality. However, the verdict on its value is still pending. We should expect some large scale evaluation of this new feature of OPACS in the foreseeable future.

REFERENCES

1. Pauline Atherton Cochrane, *Improving LCSH for Use in Online Catalogs : Exercises for Self-Help with a Selection of Background Readings*. Littleton, Colo.: Libraries Unlimited, 1986, p. 5-7.
2. *Subject Cataloging Manual: Subject Headings.* 5th ed. Washington, D.C. : Library of Congress, Cataloging Policy and Support Office, 1996. Vol. 1, section H 180, p. 1-2.
3. Alex Byrne and Mary Micco, "Improving OPAC Subject Access : The AFDA Experiment", *College & Research Libraries,* 49, no. 5 (Sept. 1988): 440.
4. Joseph R. Matthews, *Public Access to Online Catalogs*, 2nd ed. New York: Neal-Schuman, 1985, p. 8.
5. Pauline Atherton Cochrane, *Redesign of Catalogs and Indexes for Improved Online Subject Access: Selected Papers of Pauline A. Cochrane*. Phoenix, Ariz.: Oryx Press, 1985, p. viii.
6. ----- and others. *Books Are for Use: Final Report of the Subject Access Project to the Council on Library Resources.* Syracuse, N.Y.: Syracuse University, School of Information Studies, 1978.
7. Richard Van Orden, "Content-Enriched Access to Electronic Information: Summaries of Selected Research", *Library Hi Tech,* issue no. 31 (1990, no. 3): 27-32.
8. Carol A. Mandel, "Enriching the Library Catalog Record for Subject Access", *Library Resources & Technical Services*, 29, no. 1 (Jan./Mar. 1985): 5-15.
9. Byrne and Micco,op cit, p. 440.
10. Irene Wormell, "Factual Data Retrieval According to SAP Technique", *International Forum on Information and Documentation*, 8, no. 3 (1983): 15.
11. Wormell, loc.cit.,
12. Martin Dillon and Patrick Wenzel, "Retrieval Effectiveness of Enhanced Bibliographic Records", Library Hi Tech, issue no. 31 (1990, no.3): 43-46.
13. Ruth H. Makinen and Betsy Friesen, "Enhancing Online Bibliographic Records to Improve Retrieval of Reference Collection Monographs", *Bulletin of the Medical Library Association*, 83, no. 2 (Apr. 1995): 244-246.
14. *Content-Enriched and Enhanced Subject Access in USMARC Records* (MARBI Discussion Paper No. 42, rev. Feb. 7, 1991), p. 2.
15. *"Enhancing USMARC Records with Table of Contents"* (MARBI Discussion Paper No. 46) In *Advances in Online Public Access Catalogs*, edited by Marsha Ra, vol. 1, p. 105-113.
16. *Table of Contents Information* (a one-page document from LC's Electronic CIP).
17. Eduardo Peis and J. Carlos Fernandez-Molina, "Enrichment of Bibliographic Records of Online Catalogs Through OCR and SGML Technology", *Information Technology and Libraries,* 17, no. 3 (Sept. 1998): 161-172.
18. Claus Poulsen, "Tables of Contents in Library Catalogs: A Quantitative Examination of Analytic Catalogs", *Library Resources & Technical Services,* 40, no. 2 (Apr. 1996): 137.
19. Pauline Atherton and others, op. cit., p. 101.
20. *American Library Directory 1997-98*, 50thed., vol. 2. New Providence, N.J.: R.R. Bowker, 1997, p. 2353, 2467.
21. *Maclean*, Nov. 24, 1997, p. 59.
22. Walter Zimmerman's E-mail Message to The Author on Sept. 29, 1998 (Mr. Zimmerman is with Reference Services, The D.B. Weldon Library, University of Western Ontario, London, Ontario, Canada).
23. David Palmer's E-mail Message to The Author on Sept. 23, 1998 (Mr. Palmer is Librarian at

Hong Kong University).
24. Wendy Kennedy's E-mail Message to The Author on Sept. 29, 1998 (Ms. Kennedy is Head, Technical Services, University of Western Ontario, London, Ontario, Canada).