

Report on Linking Subject Headings to LC Classification Numbers and Suggestions for Automating the Classification Schedules for the Explicit Purpose of Improving Subject Access in Online Public Access Catalogs

Mary Micco

Computer Science Department, Indiana University of Pennsylvania, Indiana, PA 15705, USA

(Research supported by the Council on Library Resources and Department of Education Library Technology Grant No. R197D90020, and performed in cooperation with Carlyle Systems, Inc.)

PROBLEM STATEMENT

The average Library of Congress (LC) MARC record in today's online catalog provides a very limited number of subject access points. Certainly keywords can be extracted from the subject headings and the title but in both cases these reflect the "aboutness" of the book rather than its information content. The terms tend to be very broad and the size of the available vocabulary is very restricted. The justification for this is of course that the subject headings provide a grouping function.

CONTROLLED VOCABULARY....NOT ENOUGH HITS

We all recognize that the grouping function provided by the subject headings is needed, it could be considered essential. The problem is that the terms used often do not coincide precisely with the terms known to the hapless patron who leaves without finding the desired material. Studies by Clifford Lynch [1985] of the University of California (UC) have shown that the average result size for subject searches in UC's MELVYL catalog increased by around 60% when searches returning 0 hits were excluded from the calculation; recently for a particular week, around 40% of subject searches yielded 0 hits (Lynch, personal communication, September 26, 1990). A quite unacceptable performance... due at least in part to the fact that there is not a sufficiently large entry vocabulary.

CONTROLLED VOCABULARY...TOO MANY HITS

For a particular week, an average of 330 hits was reported for subject searches in MELVYL (Lynch, personal communication, September 26, 1990). There are some who point with pride to the fact that when the user does find the correct subject heading they will retrieve an average of 191 hits. Very few patrons are willing to deal with sets of this size. The problem is that the controlled vocabulary is sometimes too broad and therefore groups very large sets. Pointing out the need to establish new headings, Lynch [1985] listed a number of subject headings that apply to several hundred books; for example, 1550 for *Evolution*, 1561 for *Psychoanalysis*, 1571 for *Economics*. This problem is also seen in Figure 1, e.g., 19,041 hits for *Foreign relations*. If we are to improve subject access, we must find more effective means of breaking down these large clusters into more manageable subsets.

MAPPING OF NATURAL LANGUAGE TERMS

One solution to this problem, proposed by Marcia Bates [1986], is the development of a natural language mapping scheme that will provide many more access points to the record. The problem becomes how to identify and enter all the possible variants of all the terms and how to link the entry terms to the controlled vocabulary.

1. Brute Force: Entering additional keywords manually

There are librarians who have resorted to the brute force approach: boasting proudly of having entered 12 different variants of the spelling of Nietchze's name. This is completely impractical when one considers the boundless wealth of possible subjects that users may conceivably be interested in.

2. Using LCSH online for authority control

Another possibility is the use of LC Subject Headings (LCSH) online to assist users in locating the correct terms [Markey and Vizine-Goetz 1988]. Presumably such a scheme would include many more *see* and *see also* references to provide entry points into the controlled terms. While an authority control file is both useful and necessary it is a very cumbersome method of rapidly increasing the size and complexity of the entry vocabulary which will be mapped to the controlled vocabulary.

3. Automating the LC Classification Schedules (10 years away)

Karen Markey [Markey and Demeyer 1986] explored in great detail the use of information buried in the classification number. Clearly the keywords represented by the class number would provide very useful additional access points in the record. Although it should be noted for the record that these additional keywords are fairly general. While Markey addressed the use of the Dewey Decimal Classification (DDC) because it is hierarchical in nature and available in machine readable form, little has been done with the LC Classification beyond the systems analysis stage. An exception was a demonstration prototype developed by Lesk [1989] that displayed retrieved titles (from an OCLC catalog of 800,000 book citations) by either DDC or LC classes. (It also had a novel feature of plotting titles on a two-dimensional display, using DDC and LC numbers for these titles along the axes, or using as axes LC number and number of syllables in title; the difficulty is finding useful, meaningful parameters that scale neatly.) In our project we are proposing to add the keywords obtained from the captions of the class numbers to our system.

4. Adding abstracts to MARC records

A different approach is the systematic enrichment of the MARC record, with additional keywords for subject. It has been amply demonstrated in the periodical databases that the more access points you have for a record the greater are your chances of retrieving it. The database vendors accomplish the addition of keywords or subject access points by adding author-generated abstracts to all articles being indexed. Pauline Cochrane and Barbara Settel [1982] were the first to suggest that the easiest and cheapest method for books is to consider the table of contents as forming an abstract which can readily be added to the MARC record using the 690 or 653 fields. The Australian Defence Force Academy Library

The screenshot shows a search window titled 'ILSA'. The search topic is 'Korean War'. Below the search bar, it indicates 'Number Of Matches Found: 23'. There are three buttons: 'all', 'get marc', and 'get titles'. The search results are displayed in a table with four columns: the subject heading, the classification schedule, the call number, and the number of matches.

Subject Heading	Classification Schedule	Call Number	Matches
Korean War, 1950-1953	History/Asia	DS 918.15.K67	1
Korean War, 1950-1953	History/Asia	DS 918.B53	1
Korean War, 1950-1953	History/Asia	DS 918.B8	1
Korean War, 1950-1953	History/Asia	DS 918.F62	1
Korean War, 1950-1953	History/Asia	DS 918.H34	1
Korean War, 1950-1953	History/Asia	DS 918.H657	1
Korean War, 1950-1953	History/Asia	DS 918.K53	1
Korean War, 1950-1953	History/Asia	DS 918.L68	1
Korean War, 1950-1953	History/Asia	DS 918.M263	1
Korean War, 1950-1953	History/Asia	DS 919.2.A43	1
Korean War, 1950-1953	History/Asia	DS 919.2.S7	1
Korean War, 1950-1953	History/Asia	DS 919.7.A8	2
Korean War, 1950-1953	History/Asia	DS 919.K38	1
Korean War, 1950-1953	History/Asia	DS 921.T7	1
Korean War, 1950-1953	HISTORY/AMERICA(General)/U.S	E 183.8.C5	1
Korean War, 1950-1953	Armies:Organization,distribution	UA 11.5.D38	1
Korean War, 1950-1953	Armies:Organization,distribution	UA 913.K4	1

Figure 2. Subject headings connected to captions from LC Classification Schedule.

OBTAINING MEANINGFUL SUBJECT CLUSTERS

As we studied the catalog data it became apparent that each subject heading (controlled term) in the system has a potential for several if not more class numbers. So that even if the user is led through the natural language terms to the right controlled term, the material can still be scattered in a number of different topics as represented by the classification numbers assigned. In Figure 2 we have done a search for *Korean war*. Relevant material is found in a number of different places in the *DS* schedule, as well as under *E* and *UA*. Unfortunately the captions for the Library of Congress Classification Schedules are not yet available in machine-readable form, but even the very simplistic captions we provided show that the differences are not insignificant. Conversely each classification number can have numerous subject headings attached (see Figure 3).

'DD 247' History/Germany

class	subject heading	number of books
DD 247.B32	Barbie	1
DD 247.D533	Dietrich	1
DD 247.G6	Goebbels	1
DD 247.G67	Goring	1
DD 247.H5	Hindenburg	1
DD 247.H5	Hitler	2
DD 247.S375	Scholl	1
DD 247.W59	Wolff	1

Figure 3. Subject headings can be listed for each classification number.

The lack of any meaningful connection between the subject headings and the classification numbers in the system makes it very difficult for the user to pinpoint exactly the topic of interest. Besides, very few users have any idea what the class number represents and we have not provided them with even the most primitive tools to find out. We decided that in our system we would consider each unique subject heading / classification number combination as a separate subject cluster. The uncontrolled terms or natural language terms would all be mapped to the unique subject heading / class number combination (subject cluster) for that book or set of books. We had to set up rules for determining the subject cluster and finally after lengthy consultations with the Cataloging Division of the Library of Congress, we decided that the simplest and most effective way to do this was to take the classification number for the book and the first subject heading assigned to the work. This combination most closely represented the "aboutness" of the work as a whole. Many will disagree with this method of selecting a unique subject cluster but I don't think anyone can argue with the benefits to be gained.

THE ILSA SYSTEM

The experimental study was designed to develop and test an interactive library subject access system, ILSA, that would enable users to make more meaningful choices when searching for information on a topic. The software creates a dense semantic network by linking all the keywords extracted from the MARC record to the subject cluster described earlier. From this

ILSA

topic: **Suicide**

Number Of Matches Found: 49

get more

get links

LC Code	Subject Heading	LC Class	Number of References
B 1475.BB5	Ethics, Modern	PHILOSOPHY/RELIGION:History	1
BF 173.S78	Psychology, Pathological	Psychology	1
BF 441.J79	Decision-making	Psychology	1
CB 430.L54	Civilization, Modern	History of civilization	1
D 767.A.A46	World War, 1939-1945	HISTORY (General)/TOPOGRAPH	1
D 792.J3	Japan	HISTORY (General)/TOPOGRAPH	1
D 805.P7	World War, 1939-1945	HISTORY (General)/TOPOGRAPH	1
D 810.J4	Holocaust, Jewish (1939-1945)	HISTORY (General)/TOPOGRAPH	1
DD 247.G67	Goring	History/Germany	1
GN 668.R45	Women	Anthropology	1
HC 59.B25	Economic history	Economic history/conditions BY CO	1
HM 22.F8	Durkheim	Sociology (General and theoretical)	1
HM 51.J47	Sociology	Sociology (General and theoretical)	1
HM 51.S634	Sociology	Sociology (General and theoretical)	1
HN 15.E46	Social problems	Social history/Social problems/reform	1
HQ 196.P3	Prostitution	Family/Marriage/Women	1
HQ 799.A8	Juvenile delinquents	Family/Marriage/Women	1
ML 2551.A8	One hundred thousand Australian k	Literature of music	1
PK 2211.E3	Urdu poetry	Indo-Iranian languages/lit	1

Figure 4. Subject clusters related to a particular keyword.

Marc record	
CLASS NUMBER	D 792 .JJ 1984
DEWEY NUMBER	940.5426
AUTHOR	Hoyt, Edwin P., Edwin Palmer, 1923-
TITLE	The Kamikazes--Edwin P. Hoyt
PUBLICATIONS	London, Hale, 1984
SUBJECT HEADING	Japan--Nihon Kaigun Koku-bu--Shimpu Tokubetsu-Kogekitai Japan--Rikugun--Kokutai--Shimbu Tokubetsu Kogekitai World War, 1939-1945--Aerial operations, Japanese World War, 1939-1945--Campaigns--Pacific Ocean
TABLE OF CONTENTS	The Kamikazes Special attack force A-operation day B-san Iwo Jima Battle for Japan Kyushu air bases Kyushu air strikes Suicide brigade Last "Decisive battle" Okinawa Operation ten go Operation ten go II Flight of the sacred crane

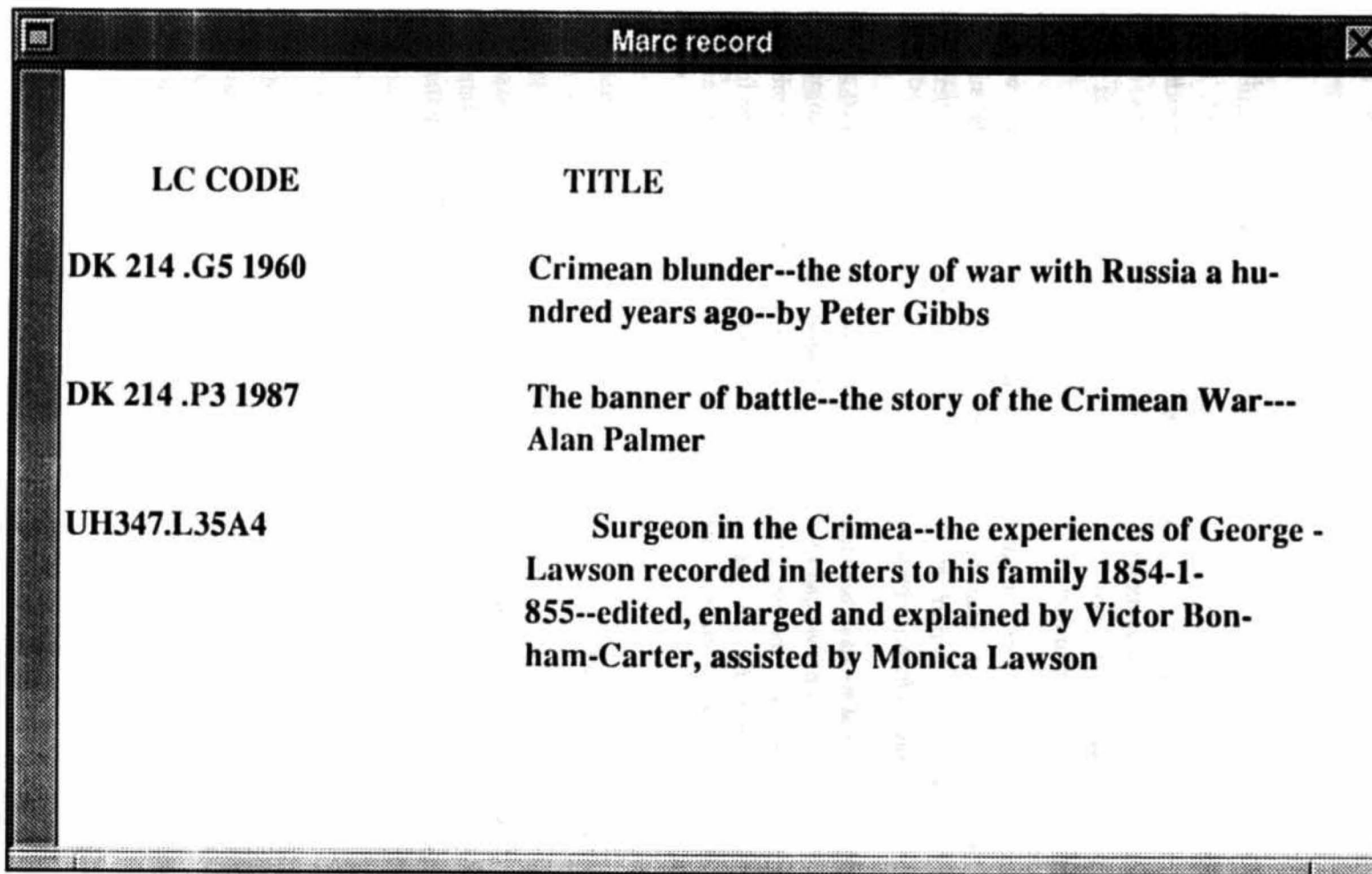
topic: **War**

Number Of Matches Found:66

all get marc get titles

Subject Heading	Number of References
Afghan Wars	1
Anglo-French War, 1793-1802	1
Austro-Prussian War, 1866	1
Chinese-Japanese War, 1894-1895	1
Crimean War, 1853-1856	3
Destroyers (Warships)	2
Falkland Island War, 1982	1
Falkland Islands War, 1982	9
First Coalition, War of the, 1792-1795	2
Franco-German War, 1870-1871	4
Indochina War, 1946-1954	1
Indochinese War, 1946-1954	2
Israel-Arab War, 1948-1949	1
Israel-Arab War, 1967	5
Israel-Arab War, 1973	3
Italo-Ethiopian War, 1935-1936	1
Korean War, 1950-1953	20
Napoleonic Wars, 1800-1814	10

Figure 5. Alternate display when number of subject clusters is excessive.



LC CODE	TITLE
DK 214 .G5 1960	Crimean blunder--the story of war with Russia a hundred years ago--by Peter Gibbs
DK 214 .P3 1987	The banner of battle--the story of the Crimean War--- Alan Palmer
UH347.L35A4	Surgeon in the Crimea--the experiences of George - Lawson recorded in letters to his family 1854-1-855--edited, enlarged and explained by Victor Bonham-Carter, assisted by Monica Lawson

Figure 6. Display of class numbers and titles for a subject cluster.

data the system can generate a variety of different responses to a request for help. As a first step when the user requests information on a topic, e.g., *suicide*, the system can quickly generate a list of class numbers and their related subject headings such as shown in Figure 4. From there the user can go directly to the book selected and view the table of contents, reconstructed from the 653 field as shown in Figure 4A.

Looking at only the list of subject headings generated in Figure 4, it is difficult to discern the differences in the meanings of the various classification numbers being reported. However if we add the captions and the hierarchy for the class numbers, we find we are presenting real choices for the users. They can look at the Psychology area, or the History of Germany, or Prostitution, or suicide as discussed in Urdu poetry. An additional benefit of our system is that we can display the counts of books in each subject cluster.

At the present time, neither readers nor librarians have this information while searching online. The only option is to scan the records themselves and to glance at the subject headings and class numbers at the bottom of each card. The user has no means of determining what subject areas the class numbers represent.

Another option that we have been able to derive from our system is that when a term produces too many links such as *war*, we can limit the search to subject headings containing the word *war* (see Figure 5). From there the user can select one, such as the one shown, *Crimean war*, which lists 3 books (see Figure 6). The user can then directly view the books in any format that they choose. All this is done by clicking with the mouse, and the user has no need to attempt to guess at the subject headings.

PROBLEMS ENCOUNTERED

It is not uncommon to find that a general classification number for a topic is then assigned *Special Topics A-Z*. If the reader lists only the classification number and its associated subject headings, it is often difficult to discern what this group is all about. For instance a search for *Fluid Dynamics* turned up class number *QA1*. A quick search of *QA1* revealed that the following subject headings among others were linked with that number:

- Algebraic topology
- Approximation theory
- Calculus,operational
- Elasticity
- Fluid dynamics
- Hilbert space
- Perturbation(mathematics)
- Sound Waves
- ...

It is not at all obvious what these subjects have in common until one discovers that the class number is for *Mathematics Periodicals.Societies...*

A second problem that occurs frequently is that within a given schedule subtle differences in emphasis may be conveyed by the class number that are lost in the subject headings: A search for

information on *Banks and Banking* may turn up the following, for example:

HG226	Money supply Banks and banking Money market
HG1394	Banks and banking
HG1586	Banks and banking
HG1616	Banks and banking Banks and banking, foreign
HG3881	Balance of payments Banks and banking, international Euro-bond market International economic relations International finance Investments, foreign

Clearly the information hidden in the classification schedules would provide meaningful data and it needs to be added to the MARC records. This proved to be much more difficult than we had anticipated. The classification schedules are not in machine-readable form, and we encountered a number of other problems. They were developed before the days of automation and no one has taken the time or trouble to adapt them for online use.

CONCLUSIONS

The concept of mapping the keywords in each record to a subject cluster representing the aboutness of the book is a powerful one that opens up many interesting possibilities. Our project has served to demonstrate that it can be fully automated. The only manual process was the keying in of the actual table of contents (about 10 minutes of work per book being added to the collection). From there the processing is merely adding a further index to the extraction routines already being run on the data. Searching the links does add some overhead but we have still obtained acceptable response times, while adding considerably to the flexibility of the system and the different browsing options we can generate. Altogether we have identified 15 different views of the records that might be useful or valuable to a searcher. I must stress that this is research in progress and the final analysis of the data has not yet been completed. We are still in the process of designing the user interface.

This research is being supported by the Council on Library Resources and the Department of Education through a Library Technology grant. The OPAC software is provided by Carlyle Systems, Inc. We selected the NeXT machine as the hardware platform with the Unix operating system (Mach) and Sybase as the relational database. We are taking advantage of NeXTStep to provide a Hypertext like interface.

REFERENCES

Alberico, Ralph and Mary Micco. *Expert systems for reference and information retrieval*. Westport, CT: Meckler Publishing Co., 1990.

120 PROCEEDINGS OF THE 1ST ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

Bates, Marcia J. "Subject access in online catalogs: a design model." *Journal of the American Society for Information Science* 1986 Jan;37(6):357-376.

Lesk, Michael. "What To Do When There's Too Much Information." *Hypertext '89 Proceedings, November 5-8, Pittsburgh, Pennsylvania, Special Issue — SIGCHI Bulletin*, November 1989. New York: Association for Computing Machinery, 1989. pp. 305-318.

Lynch, Clifford A. "Cataloging Practices and the Online Catalog." *ASIS '85, Proceedings of the 48th ASIS Annual Meeting, Las Vegas, Nevada, October 20-24, 1985*. pp. 111-114.

Markey, Karen and Diane Vizine-Goetz. "Increasing the accessibility of Library of Congress subject headings in online bibliographic systems." *Annual Review of OCLC Research July 1987-June 1988* Dublin, OH: OCLC, 1988. pp. 32-34.

Markey, Karen and Anh Demeyer. *Dewey Decimal Classification Online Project: Evaluation of a Library Schedule and Index Integrated into the Subject Searching Capabilities of an Online Catalog*. OCLC Research Report Series, no. OCLC/OPR/RR-86/1. Dublin, OH: OCLC, 1986.

Settel, Barbara and Pauline A. Cochrane. "Augmenting subject descriptions for books in online catalogs." *Database* 1982 Dec;5(4):29-37.