Redundancy and Uniqueness of Subject Access Points in Online Catalogs

Hong Xu and F. W. Lancaster

An analysis of 205 records selected at random from the OCLC Online Computer Library Center, Inc. Online Union Catalog showed considerable overlap (duplication) among the subject access points provided by the title, subject heading, and classification number fields. Little more than four unique (unduplicated) access points were found, on average, per record. While title and class number fields do add some access points not provided by subject headings, the increase is less than many librarians might have expected. It is suggested that the online catalog might outperform the card catalog more in precision than in recall.

Card catalogs, as implemented in the majority of libraries, provided very limited subject access possibilities. In the United States, they provided subject access only by means of a very small number of subject headings. In fact, based on a sample of more than 50,000 monographic records from the OCLC Online Computer Library Center, Inc., Online Union Catalog (OLUC), O'Neill and Aluri (1981) reported that, on the average, there were only 1.41 subject heading/subheading combinations per record and only 1.32 unique subject headings per record.

It is generally assumed that online catalogs have greatly improved subject searching capabilities. Even without adding to the conventional record (e.g., terms from contents pages or other sources), the number of subject access points (SAPs) was increased merely by making other fields searchable. (In this paper, the term "subject access point" refers to any element in a bibliographic record that is indicative of the subject of the item represented: a subject heading, classification number, or words appearing in titles, subject headings, or elsewhere.) Obviously, the title field and the classification number field both include elements that might be useful in subject searches. For some items, however, it is possible that title and classification number add no SAPs that are not already provided by the subject heading. To take an extreme hypothetical example, if a book entitled *Birds* has a single subject heading BIRDS and the Dewey class number 598, the title and classification number fields give no SAPs not already provided by the subject heading.

In an earlier study, Lancaster et al (1991) suggested that title and class numbers seem frequently to duplicate

Hong Xu (hgxu@sis.pitt.edu) is Assistant Professor, Department of Library and Information Science, University of Pittsburgh. F. W. Lancaster (gslis@alexia.lis.uiuc.edu) is Professor Emeritus, Graduate School of Library and Information Science, University of Illinois at Urbana–Champaign. Manuscript received June 3, 1997; accepted for publication August 18, 1997.

the subject headings assigned to a book rather than provide further access points. This was the motivation for the present investigation.

The study was performed to determine to what extent titles and classification numbers provide SAPs not already provided by subject headings in a typical catalog record. The principal hypotheses guiding the study were:

1. In typical catalog records, the SAPs provided by classification number (CN), title (TI), and subject heading (SH) tend to overlap (duplicate) each other considerably.
2. In typical catalog records, the CN, TI, and SH fields are significantly different with respect to the number of *unique* SAPs they provide. *Unique* is defined here as occurring in one field but not the other two.

Method

The study was performed on a sample of records drawn from the OLUC in *Dewey Decimal Classification* (*DDC*) classes 300 (Social Sciences), 500 (Natural Sciences/Mathematics), 600 (Technology), and 700 (the Arts). A 3x4 level factorial design was selected to determine the level of duplication among the SAPs in different subject areas. This is appropriate because of the involvement of three subject fields (title, subject heading, class number) and four subject areas (classes 300, 500, 600, and 700). Based on Cohen (1988) and Stevens (1990), for a 3x4 analysis of variance (ANOVA) involving repeated measures, the sample size can be quite small: 44 records in each of the four main cells, or a total of 176 records.

As of August 10, 1994, the OLUC contained more than 734,000 records that satisfied the requirements for the study (i.e., referred to monographs in the four Dewey classes selected). To allow for the fact that some records drawn at random might not be suitable for use in the study, a somewhat expanded stratified sample (i.e., proportional to the number of records in each class) was drawn to ensure that each of the four classes was represented with at least the minimum number of records (i.e., 44) needed. After discarding records outside the parameters set for the study (e.g., books published before 1990, books without subject headings, books in languages other than English), the final sample consisted of 205 records: 58 in *DDC* class 300, 46 in class 500, 46 in class 600, and 55 in class 700. Books in non-English languages were omitted to simplify the analysis. The date restriction (1990–94 publication date) was imposed to avoid the possibility of encountering significant variations in cataloging policy, including widely different subject headings or classification numbers, over a considerable period of time.

Testing of the hypotheses involved a comparison of the contents of the title, subject heading, and classification number fields in the sample records. Within the title field (MARC 245), other title information was also considered. In the subject heading field (MARC 6xx), subheadings were considered as well as main headings, with the exception of subheadings indicating type or form of publication rather than subject. Dealing with the classification field (MARC 082) was more complicated because class numbers had to be translated into words (e.g., 327.73 translates into "foreign relations" and "United States") by use of *DDC* captions, notes, and auxiliary tables.

Unlike some earlier authors (e.g., Markey and Calhoun 1987; Frost 1989), we dealt with comparisons at the "idea" level rather than the word level. For example, "foreign policy," "foreign affairs," and "international relations," when associated with "United States," are so closely interrelated that it would be virtually impossible to distinguish them. For all practical purposes, then, we considered them to be synonymous.

Dealing with SAPs at the idea level clearly involves some subjectivity. Because, for practical reasons, all decisions on equivalency were to be made by Xu, it was necessary to determine whether or not her decisions were likely to be supported by the majority of others. A validation procedure was used for this purpose.

The validation was based on 30 sample records for which Xu had already made

TABLE 1
RULES USED FOR DETERMINING EQUIVALENCY AMONG SUBJECT ACCESS POINTS

1. Ignore words that reflect method of presentation rather than subject content (e.g., "report," "approach," "workshop").
2. In proceedings of conferences, or other meetings, consider only words indicative of subject covered. Ignore words indicative of conference location or frequency
3. A specific subject idea (species) includes the idea immediately more general (genus).
   Example: a title includes the word "education" but the subject heading specifies "primary education." In this case, both title and subject heading cover the "education" idea, but the subject heading introduces the additional idea of "primary" (i.e., a particular age group).
4. Words or phrases associated with the subject access points are considered synonymous (and therefore equivalent) under the following conditions:
   a) absolute identity (e.g., subject heading = birds, titles = birds)
   b) abbreviations (e.g., USA = United States)
   c) popular versus more "serious" usage (e.g., stamp collecting = philately)
   d) equivalency at phrase level although not at word level (e.g., church music = sacred music = liturgical music, even though "church," "sacred," and "liturgical" are not synonymous)
   e) Implicit synonymy (e.g., "America" in title is equivalent to "USA" in subject heading field when it is clear that author is using "America" as shorthand for USA)
   f) singular/plural equivalency (e.g., mouse = mice)
   g) standard versus slang usage (e.g., high fidelity = hi fi
   h) variant spellings (e.g., catalog = catalogue, online = on-line)
   i) variant usage within English (e.g., railroad = railway)
   j) words with the same etymological root (e.g., electroluminescence = electroluminescent,Egypt = Egyptian)
   k) different expressions of the same historical period (e.g., 1930's = 1930–1939 = "the thirties")

decisions on equivalency and uniqueness of SAPs for the three fields. Thirty masters students from the Graduate School of Library and Information Science, University of Illinois at Urbana–Champaign, participated voluntarily in the validation. The 30 sample records and the 30 students were randomly grouped into six units such that each student examined five records and each record was examined by five students.

For each item, a student was given (a) the contents of the three fields as they appeared in the OCLC record, (b) Xu's decisions on equivalency or uniqueness of access points within these fields, and (c) a set of rules followed by Xu in determining equivalency (see table 1). Students were to indicate whether or not they agreed with the decisions and to explain reasons for any disagreement.

For each group of five records, then, it was possible to determine to what extent the group of students agreed with Xu's decisions. For example, for one group of five records, she had identified 27 unique SAPs. All five students agreed with 21 of these, and four of the five students agreed with the other six (agreement means that the student agreed that the SAP was not equivalent to any of the other SAPs in the record in which it appeared). It was not to be expected that all students would agree with all of Xu's decisions. However, about 78% of her decisions were supported at the 100% level (all five students) and an additional 18% at the 80% level (four out of five students). This level of agreement was considered to be sufficiently supportive to allow Xu to proceed with the other decisions without further validation or corroboration of this type, especially because many of the disagreements that did occur stemmed from a misinterpretation of the rules established by Xu (e.g., a general idea was not considered unique when it is included, explicitly or implicitly, in a more specific idea: "architecture" is included in "domestic architecture," "cyanides" is included in "sodium cyanide").

In fact, Xu's decisions, while ultimately subjective, were always supported by the use of appropriate reference tools (encyclopedias, dictionaries, thesauri, glossaries) or, if necessary, by consulting others on campus

TABLE 2
EXAMPLE OF RESULTS OBTAINED FROM ONE OF THE SAMPLE RECORDS

Title:	Efficient masonry housebuilding
Subject headings:	Masonry—Great Britain House construction—Great Britain
Class number:	693, construction in specific types of materials and for specific purposes
Unique subject ideas:	Masonry
	Houses
	Construction
	Great Britain
Subject ideas represented in title:	masonry, houses, construction (= building)
Subject ideas represented in subject headings:	all four
Subject ideas represented in class number	construction (only)

more familiar with the subject matter.

It should be noted that equivalence among SAPs was determined by the context of the terms within the bibliographic record itself, not the context (e.g., classification schedules) from which the terms were drawn. Thus, U.S. + foreign relations, U.S. + international relations, and U.S. + foreign policy were considered so close conceptually that we felt unable to distinguish clearly among them. Put differently, it was felt that a book dealing with U.S. foreign policy must, in some sense, deal with its international (foreign) relations and vice versa, despite the fact that these terms might be treated somewhat differently in the classification schedules, the list of subject headings, or both.

RESULTS

The types of results obtained in the study are illustrated in the example given in table 2. In this example, four different SAPs have been identified. Of these, the subject heading field includes all four, the title field three, and the class number field only one.

The major results of the study, at the macro level, are shown as a Venn diagram in figure 1. Because the comparison across subject areas is not the main focus of this paper, these results are not presented here. Altogether, 844 unique SAPs were assigned to the 205 items, the average per record being 4.12. Of these, 210 (24.88%) were duplicated in all three fields. At the other end of the spectrum, 414 (49.05%) of the total headings appeared in only one field: 209 (24.76%) appeared only in the subject heading field; 118 (13.98%) appeared only in the title field; and 87 (10.31%) appeared only in the class number field.

As figure 1 shows, the subject heading (SH) field alone contributed 634 (75.12%) of the SAPs, while the title (TI) field alone contributed 458 (54.27%), and the classification number (CN) 406 (48.10%).

Another way of looking at these data is in terms of the proportion of unique SAPs contributed by each field. As mentioned, the total number of SAPs unique to a single field was 414 (209+118+87) and, of these, the SH field contributed 209 (50.48%), the TI field 118 (28.50%), and the CN field 87 (21.02%).

As figure 1 indicates, the greatest overlap occurred between the SH and TI fields, with slightly less between the SH and CN fields. The least overlap occurred between the TI and CN fields largely because titles tended to provide much greater specificity.

Table 3 presents the results of a one-way ANOVA for overlapping SAPs between two fields. Because the average

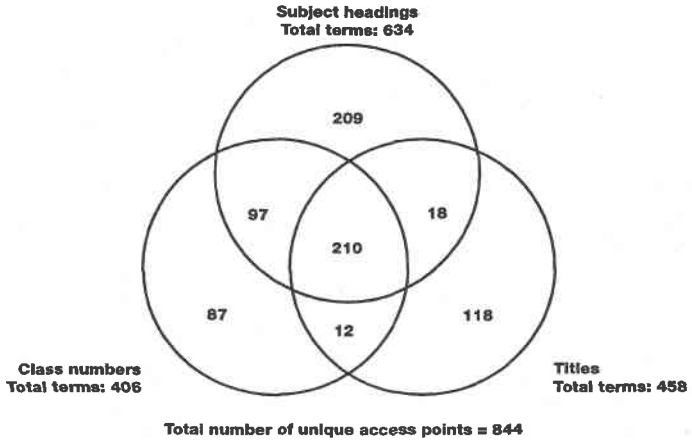

Figure 1. Overlap of Subject Points in Three Subject Fields.

number of overlapping SAPs between every pair of fields is significantly different ($F=38.30$, $p<0.0001$), the null hypothesis associated with our first hypothesis—that there are no significant differences in the average number of overlapping SAPs among each pair of fields—is rejected.

Hypothesis 2—that the three fields are significantly different with respect to the number of unique SAPs they contribute—was tested by comparing the unique SAPs existing in the three fields, taking all three together and looking at each pair separately.

Table 4 shows the results of a one-way ANOVA for the unique SAPs provided by the three fields: the three fields provided a significantly different number of SAPs ($F=24.29$, $p<0.0001$) so the null hypothesis—that there are no significant differences among the three fields in the number of unique SAPs they provide—is rejected. In other words, we can conclude that the three fields are significantly different with respect to the number of unique (nonoverlapping) SAPs they contribute.

The pairwise comparisons (table 5) indicate that each field is significantly different from each of the other two fields in

TABLE 3
ONE-WAY ANOVA FOR OVERLAPPING SAPS BETWEEN TWO FIELDS

Sources	Degrees of Freedom	Sum of Squares	Mean of Squares	F-value
c	2	29.08	14.54	38.30*
Error	402	152.60	0.38	

Notes: * p 0.0001

TABLE 4
ONE-WAY ANOVA FOR UNIQUE SUBJECT ACCESS POINTS PROVIDED BY THREE SUBJECT FIELDS

Sources	Degrees of Freedom	Sum of Squares	Mean of Squares	F-value
Field	2	42.13	21.07	24.29*
Error	402	348.67	0.87	

Notes: * p 0.0001

the number of unique SAPs provided.

CONCLUSIONS

Remarkably few unique SAPs exist in a typical online public access catalog record—little more than four (actually 4.12 in our sample). In this study, it has been shown that online catalogs offer less improvement over card catalogs in the number of unique SAPs provided than many librarians might have expected, at least for the four subject areas covered. If the number of unique SAPs per record can be considered a measure of "retrievability," titles add only modestly to subject headings alone, and classification numbers contribute very few access points not provided by the other fields. In subject access, then, the main advantage of the online catalog over the card catalog might lie more in allowing greater discrimination in searching (terms from different fields can be combined; titles offer greater specificity; searches can sometimes be restricted by date, language, or other criteria) than in providing a more complete (exhaustive) representation of the subject matter of each book. Put differently, the potential improvement in precision might be greater than the potential improvement in recall.

WORKS CITED

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Frost, C. O. 1989. Title words as entry vocabulary to *LCSH:* Correlation between assigned *LCSH* terms and derived terms from titles in bibliographic records with implications for subject access in online catalogs. *Cataloging and classification quarterly* 10, nos. 1 & 2: 165–79.

Lancaster, F. W., et al. 1991. Identifying barriers to effective subject access in library catalogs. *Library resources & technical services* 35: 377–91.

Markey, K., and K. Calhoun, 1987. Unique words contributed by MARC records with summary and/or contents notes. *Proceedings of the American Society for Information Science* 24: 153–62.

O'Neill, E. T., and R. Aluri, 1981. Library of Congress subject heading patterns in OCLC monographic records. *Library resources & technical services* 25: 63–80.

Stevens, J. 1990. *Intermediate statistics: A modern approach.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

TABLE 5
PAIRWISE COMPARISONS FOR UNIQUE SAPS IN THREE FIELDS

Fields	Significance Level
CN–TI	p< 0.05
CN–SH	p<0.0001
TI–SH	p<0.0001