Pergamon

Information Processing & Management, Vol. 30, No. 3, pp. 327-342, 1994
Copyright © 1994 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0306-4573/94 $6.00 + .00

0306-4573(93)E0015-G

INDEXING TERMS IN THE LISA DATABASE ON CD-ROM

WILLIAM HOOD and CONCEPCIÓN S. WILSON
School of Information, Library and Archive Studies, University of New South Wales,
P.O. Box 1, Kensington, NSW, 2033 Australia

(*Received* 19 *November* 1992; *accepted in final form* 20 *May* 1993)

Abstract – This paper summarizes the findings of a recent study on the indexing practices used in the Library and Information Science Abstracts (LISA) database. The indexing terms (DE), the date each record was added to the file (DA), the Accession Number (AN), and the Classification Code (CC) of each record were extracted from the complete CD-ROM database. Adjustments to standardize the DE terms were made, and the adjusted data set was analyzed for average number of headings per record per DA year, the rank frequency and rank size distribution of DE classes, and the frequency distribution of the number of DEs per record per DA. The results show that a large number of headings are used once or twice over the whole database. The years in which DE terms first appeared was analyzed. A comparison of the use of a particular Classification Code with the DE terms in each record was also undertaken. There was a strong but not total match between the CCs and DEs used. Some attention is given to the chain indexing procedure used by LISA to account for the pattern of DE usage. The concluding section looks at scope for further research on LISA and other databases.

1. INTRODUCTION

This study analyzes the indexing terms used in the *Library and Information Science Abstracts* (LISA) database. There were two broad reasons for undertaking this study: first, to try to understand more about the indexing used in LISA. This is important because LISA is the main bibliographic database in the field of Library and Information Science, and understanding how the index is constructed would be beneficial for anyone in the field who uses this database for teaching and research. The second reason is to try to develop some measures that could be applied to other databases in evaluating and comparing the indexing processes used in different databases.

The literature on the analysis, evaluation, and comparison of the LISA database generally falls into the following categories:

1. Evaluation of a particular database on a particular medium such as LISA on CD-ROM. Evaluations in this category address such issues as installation, searching capabilities, ease of use, documentation, initial and updating cost, output options, and indexing features (Day, 1988; Hartley, 1989; Moore, 1988a; Terbille, 1990).

2. Historical perspective of a particular database showing its growth, development, and future directions. This category includes papers that outline the origins, working methods, subject coverage, currency, bias, indexing, and classification guidelines of a particular database such as LISA. The authors of such papers are generally present or past editors of the database in question (Edwards, 1976; Tomlinson, 1986).

3. Description of specialized indexing methods used in particular databases. As in category 2, these papers are written by persons closely associated with the database, such as directors or editors. Topics covered in these papers include detailed description of indexing procedures, classification guidelines used, and numerous examples from the database to illustrate the procedures used (Edwards, 1975; Moore, 1988b).

Correspondence and requests for reprints should be addressed to William Hood, School of Information, Library and Archive Studies, University of New South Wales, P.O. Box 1, Kensington, NSW, 2033 Australia.

4. Comparison of various characteristics of several databases on a particular subject area. These papers are often more than just a description of the databases under discussion; they often are based on experimental data used in measuring such aspects as the quality of indexing, the number of terms per document or concepts per document, the effectiveness of search vocabularies across the various databases, overlaps among databases, and authorship patterns (Bottle & Efthimiadis, 1984; Chu & Ajiferuke, 1989; Feng, 1989; Sievert & Verbeck, 1987).

With the exception of historical perspective papers, all the studies above have attempted to analyze, describe, or compare using only a sample of the database or databases. With the advent of databases on CD-ROM, it is now possible to examine in great detail aspects of an *entire* database and not just a portion of it, based on whatever sampling technique is used. This study differs from earlier studies in that it looks at all of the 111453 records of the 1969 to September 1991 edition of LISA on CD-ROM (LISA, 1991). It is hoped that in examining various aspects of the indexing terms used for the entire database, our analysis will be able to offer yet another method of describing and measuring the quality of indexing used in databases such as LISA.

Finally, we wish to note that no effort was made to review previous studies on topics relating to the use of controlled vocabulary in databases, such as the frequency of assignment of index terms. For comparisons of certain aspects of this study to earlier works, the reader is directed to Lancaster's (1972) comprehensive review of all aspects of the use of controlled vocabulary in information retrieval systems—for both information storage (index term assignment) and information retrieval (thesaurus use in subject searching).

2. LISA'S SUBJECT INDEXING SYSTEM

LISA uses the chain indexing technique for compiling the alphabetical subject index entries in the printed product and the list of index terms (DE) in the electronic products. Prior to invoking the chain indexing procedure, LISA indexers compile specific subject phrases in words that are tailor-made to fit the abstracts (Edwards, 1975). The indexers then use a faceted classification scheme to translate the subject phrases into a classification number or code (CC). The classification scheme used by LISA is one where only single concepts are listed in the classification schedule, and multi-concepts or compound subjects are classified by synthesizing, or number building, the notations from the appropriate parts of the schedules. The notations are combined by strict adherence to a well established sequence of facets—the citation or preferred order. Thus, the notations or classification numbers (CC) in LISA are derived from the feature headings (FH) or the verbal statements of the subject. Feature headings for different documents generally are not identical with each corresponding classification code. The classification codes (CC) are then used to start the chain indexing procedure from a magnetic disc Classification File at the computer bureau that LISA uses. The following example illustrates this procedure.

> A document on 'Centralized on-line cataloguing in university libraries in the Netherlands' would be classified at TogsNjrGdD492. The computer takes the first character of the notation and matches it against the Classification File. The process is then repeated for the first 2, first 3, . . ., etc. characters of the notation, i.e.:
>
> T
> To
> Tog
> Togs
> TogsN
> TogsNj
> TogsNjr
> TogsNjrG
> TogsNjrGd
> TogsNjrGdD
> TogsNjrGdD4
> TogsNjrGdD49
> TogsNjrGdD492
>
> . . . each chain indexing string drawn from the Classification File is also matched against a Cross-Reference File and draws any relevant *see* or *see also* references (Moore, 1988b, p. 14).

This automatic process results in a list of index terms (DE) for each document in the electronic products and an updating of the printed alphabetical subject index to include newly indexed documents. The feature heading (FH) is thus coextensive with the scope of the document as expressed by the index terms (DE), whereas the classification code (CC) is generally broader in subject scope.

The top terms in a hierarchy using the chain indexing procedure cannot be seen as 'true' descriptors. Descriptors (DE) are specific to the scope of a document. In the example above, some of the index terms that the first character, T, in the classification code drew from the Classification File are: Technical processes and services, Library materials, Information storage and retrieval, etc. These terms should be viewed as broader terms of the document, which have resulted from the 'upposting' procedure inherent in chain indexing. Not until the characters TogsNjr are read by the computer does the specific scope of the document begin to emerge—Centralized online cataloguing.

3. INITIAL ANALYSIS

The analysis for this study was based initially on the three fields AN (Accession Number), DA (Date Added), and DE (Descriptor or Indexing term) of each of the records of the LISA database (Hood & Wilson, 1992).* A number of discoveries were made in this stage. The CC (Classification Code) field was later examined in conjunction with the DE field.

Errors were discovered in the DA and AN fields, such as the repetition of the information in the field. With regard to the DE terms, many were found to be repeated in the same record. Some records even had the same DE appearing six times. DE terms were also found to terminate either with a dash or not. A pattern emerged whereby single word DE terms almost always terminated with a dash prior to DA = 1989. With hyphenated, multiple-word DE terms, the general pattern (with some exceptions) found was that in DA = 1975, the multiple word DE terms did terminate with a dash, whereas in all other years, they did not. Note that in the examples given in Table 1, LISA does not differentiate between hyphens and terminating dashes.

Other inconsistencies found in the DE terms are categorized in Table 2.

4. SEARCHING IMPLICATIONS

Some of the inconsistencies discussed above have serious implications for searching LISA on CD-ROM; others, though not impeding search facilities, do represent a certain lack of control in indexing practice (as well as appearing inelegant).

DE inconsistencies with no searching implications:

- duplicate DE terms in the same record;
- DE terms with quotation marks (quotation marks are removed in the inverted index, as well as from a user's search term);
- case discrepancies in DE terms.

*A full report of this study also included an analysis of the PY (Publication Year) field.

Table 1. Use of terminating dashes in DE terms

DE type	General pattern found	Examples
Single word	< 1989	Acquisitions-; AACR-
	>= 1989	Acquisitions; AACR
Hyphenated multiple word	1975	Chinese-People's-Republic-; Abstracting-services-
	not 1975	Chinese-People's-Republic; Abstracting-services

Table 2. Inconsistencies in DE terms

Inconsistencies in DE terms	Examples
DE terms differ by the case used	Aabenraa-Denmark
	aabenraa-Denmark
DE terms differ by the presence or absence of parenthesis	Adelaide-South-Australia
	Adelaide-(South-Australia)
DE terms differ in use of periods	USA
	U.S.A.
	U.S.A
DE terms differ in use of quotation marks	'Alternative'-libraries
	Alternative-libraries
DE terms do not adhere to controlled vocabulary practice; as the same term is expressed in a variety of formats	Acquired-Immune-Deficiency-Syndrome-(AIDS)
	Acquired-Immune-Deficiency-Syndrome-AIDS
	AIDS-(Acquired-Immune-Deficiency-Syndrome
DE terms misspelled or outdated	AIDS-(Acquired-Immune-Deficiency-Syndrome)
	AIDS-(Aquired-Immune-Deficiency-Syndrome)
	AIDS-Acquired-Immune-Deficiency-Syndrome
	AIDS-Auto-Immune-Deficiency-Syndrome

DE inconsistencies with searching implications:

- presence or absence of trailing dashes;
- single word DE terms: DE terms with trailing dashes are indexed both with and without the dashes, however terms without trailing dashes are only indexed as is. The only way to be sure to get all records using such a term is to search the DE without using the trailing dash. However, this method will also retrieve multiple-word terms containing this word;*
- multiple word DE terms: Each multiple word DE is indexed both as a phrase and as single words. One way to be sure to get both forms of the DE is not to use a trailing dash, but to truncate after the last alphanumeric character. The other way to achieve this would be to "OR" the DE term with and without the trailing dash;
- DE terms with parentheses: These terms are indexed as two separate phrases as well as word indexed. There is no index entry for the whole term (e.g., "ADELAIDE-(SOUTH-AUSTRALIA)" is indexed under ADELAIDE; SOUTH-AUSTRALIA; SOUTH; AUSTRALIA; thus to retrieve this term, a searcher would have to "AND" the two parts of the term together: "ADELAIDE AND SOUTH-AUSTRALIA").

5. ANALYSIS AFTER ADJUSTMENTS

Due to the discrepancies found in the DE terms, the following steps were taken to adjust the terms in the remaining processing:

- Trailing dashes were removed.
- All terms were converted to upper case.
- All terms duplicated in the same record were converted to one heading.
- Headings designated only with "-" were converted to "no heading."

Removal of trailing dashes and conversion to upper case were done to standardize the DE terms that would otherwise appear as different terms. Removal of repeated DEs was

*This is not a problem in the DIALOG online version of LISA, which has the /DF (or descriptor full) qualifier, which permits the retrieval of single word DE terms, without the retrieval of multiple word DE terms containing that word.

done to gain a more realistic picture of the number of DE terms assigned per record. Terms designated with a "-" were converted to no term, to accurately determine DE counts. Further adjustments, such as the removal of other punctuation or spelling corrections, were not undertaken at this point.

Table 3 shows the results of the adjustments.

5.1 *Average number of DE terms per record*

The average number of adjusted DE terms was calculated for each DA. The overall trend of the number of terms used per record is shown in Fig. 1 and in the Appendix. During 1977–78, LISA was assigning more DE per record than in earlier or later years. Since 1979, LISA has consistently been assigning an average of about six to seven DE terms per record. Though not explicitly stated by LISA editor Moore (1988b), one can conclude that the investigation of LISA's production and indexing costs in 1978 has resulted in the optimal average number of DE per record. One can only speculate as to how the optimum number of DE per record was derived; however, it is hoped that the optimum was derived by having balanced the indexing and storage costs with those of subject-retrieval effectiveness.

5.2 *Distribution of the frequency of occurrences of DE terms*

The distribution of the frequency of occurrences of DE terms was analyzed and presented in Table 4 and Figs. 2 and 3. Various aspects of the top 10 and the bottom 10 of the 506 equal productivity classes of DE terms are shown. The last two columns of Table 4 list values only for classes #1 to #10, where the number of DE terms per class equals one. Classes #497 to #506 have large numbers of DE terms per class, and the percentage of the total records (111453) having these DE terms can only be expressed as a range.

According to Nelson and Tague (1985), there are two types of distributions commonly used to summarize word frequency data: the *rank-frequency* and the *frequency-size*. In this study the concentration of high-frequency terms (e.g., TECHNICAL-PROCESSES-AND-SERVICES was used 25942 times in 23.3% of the total records in the database) is best shown by the Zipf or rank-frequency plot, and the scatter of low-frequency terms (e.g., 18087 or 64.2% of all DEs were used only once) is presented in a Lotka frequency-size plot.

The Zipf rank-frequency distribution plots the number of occurrences per DE in each class (Column 2) by the cumulation of the number of DEs per class in class order (Column 5). The maximum DE rank in each class is in order of decreasing productivity or occurrences of DE in each class with no ties in the ranks. For example, class #506 has ranks 10105 to the maximum of 28191.

Regression lines were fitted to, and correlation coefficients and slopes calculated for three parts of the Zipf plot:

1. Part 1 includes DE terms with ranks 1 to 10, and has a negative correlation coefficient of −0.930 and a slope of −0.413.
2. Part 2 includes DE terms with ranks 11 to 273 (the natural splitting point between high- and low-occurring DE terms). The negative correlation coefficient is −0.997 and the slope is −1.05. DE ranks 1 to 273 constitute approximately 1% of the DE terms in the high-frequency portion of the plot. This split corresponds to that found in the empirical data tested by Nelson and Tague (1985).
3. Part 3 includes DE terms with ranks 731 to 6515, and has a negative correlation coefficient of −0.999 and a slope of −1.53.

Table 3. Number of DE terms before and after adjustment

	Number of unique DEs	Number of total DEs
Non adjusted	31068	693888
Adjusted	28191	669403

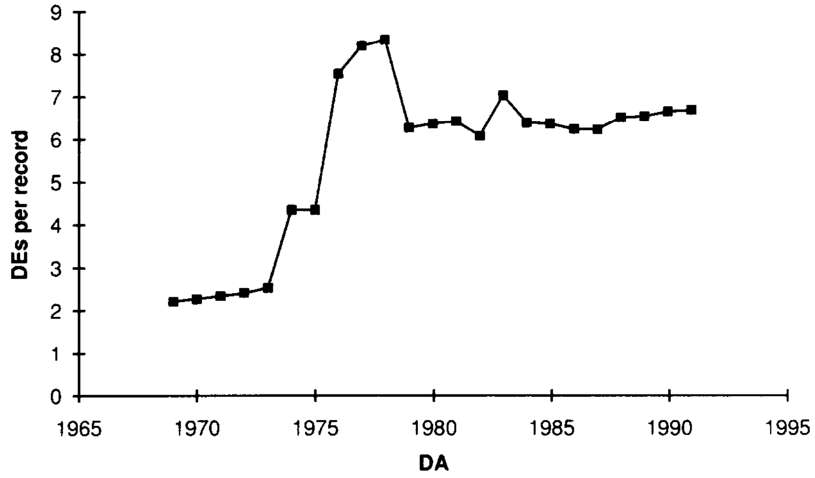

Fig. 1. Average number of adjusted DE terms per record by DA.

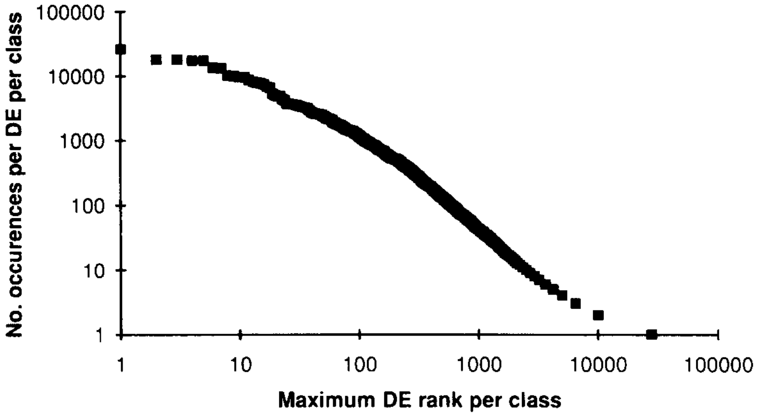

Fig. 2. Zipf Rank-Frequency plot: Number of occurrences per DE in class versus maximum DE rank in class.

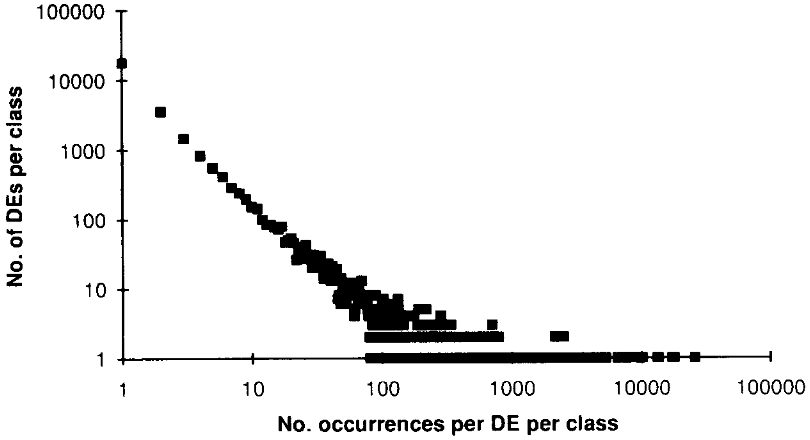

Fig. 3. Lotka Frequency-Size plot: Number of DEs in class versus number of occurrences per DE in class.

Table 4. Distribution of the frequency of occurrences of DE grouped into equal frequency of occurrences classes

Equal product Class #	No. occurrences per DE in class	No. DE in class	% Total no. DE in class	Max. DE rank in class	Top ten DEs	% of 111453 records with DE
1	25942	1	0.0	1	TECHNICAL-PROCESSES-AND-SERVICES	23.3
2	17922	1	0.0	2	INFORMATION-STORAGE-AND-RETRIEVAL	16.1
3	17919	1	0.0	3	INFORMATION-RETRIEVAL	16.1
4	17361	1	0.0	4	READER-SERVICES	15.6
5	17347	1	0.0	5	SERVICES	15.6
6	13512	1	0.0	6	SUBJECT-INDEXING	12.1
7	13313	1	0.0	7	INFORMATION-WORK	11.9
8	10161	1	0.0	8	COMPUTERISED-INFORMATION-STORAGE-AND-RETRIEVAL	9.1
9	9825	1	0.0	9	PUBLIC-LIBRARIES	8.8
10	9569	1	0.0	10	STOCK	8.6
...	...	...	...	...		
...	...	...	...	...		
497	10	153	0.5	2516		
498	9	198	0.7	2714		
499	8	242	0.9	2956		
500	7	289	1.0	3245		
501	6	413	1.5	3658		
502	5	561	2.0	4219		
503	4	836	3.0	5055		
504	3	1460	5.2	6515		
505	2	3589	12.7	10104		
506	1	18087	64.2	28191		

The Lotka frequency-size distribution plots the number of DEs per class (Column 3) by the number of occurrences (or productivity) per DE in each class (Column 2). The Lotka plot as shown in Fig. 3 is in reverse order starting with Class #506 (with a productivity or occurrence of 1 and the number of DEs per class of 18087) to Class #1 (with a productivity of 25942 and only 1 DE in the class). The Lotka plot distinguishes low-productivity DE classes, whereas the Zipf plot distinguishes high-productivity DE classes.

A regression line was fitted to, and the correlation coefficient and slope calculated for the low number of occurrences (from 1 to 70) per DE class. The negative correlation coefficient is −0.984 and the slope is −1.73.

The Pratt (1977) Index, a bibliometric measure of class concentration, is 0.936. The Gini (1909) Index, an econometric measure of income concentration nearly identical to the Pratt Index, is similarly 0.936. In both measures, "0" indicates equal distribution over class and "1" indicates total concentration in one class.

5.3 *DE terms used only once or twice in the whole database*

The large number of DE terms used only once or twice was analyzed to see the distribution by year added (DA) to the file, as shown in the Appendix and Fig. 4.

The percentages of DEs used once or twice point to the sharp decline between 1978 and 1979. For DEs used once, the drop was from 5.1% to 1.2%; for DEs used twice, it dropped from 2.1% to 0.6%. This dramatic drop is most likely related to the LISA inves-

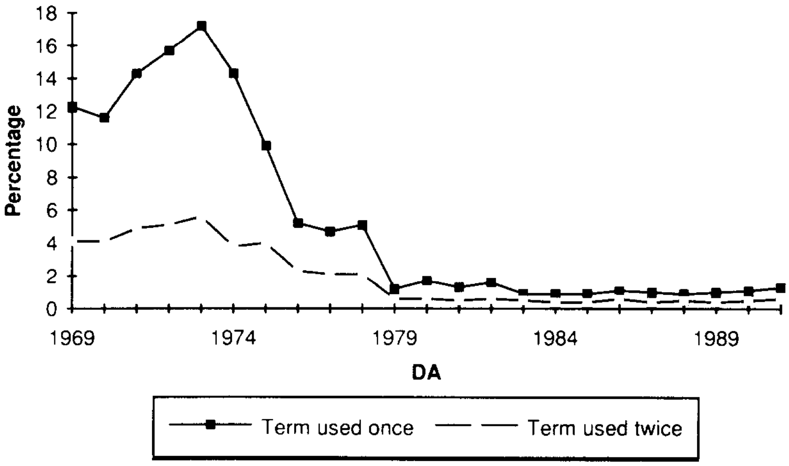

Fig. 4. Percentages of DE terms used once or twice by DA.

tigation and subsequent changes of 1978/1979. "From January 1979, large reductions in the physical size of the indexes were achieved . . ." (Moore, 1988b, p. 13).

5.4 *Frequency distribution of number of DE terms per record*

The frequency distribution of the number of DE terms per record (after adjustments) is shown in Fig. 5.

Because of the skewed distribution of the number of terms per record, the following statistics are given. The *mean* number of terms per record is 6.0; the *median* value is 5, and the *mode* is 3. From 1985 to the present, only two records were assigned 20 DEs; the rest were assigned 19 or fewer (Hood & Wilson, 1992). Moore's (1988b) investigation and resulting cost-reduction program is evident in the number of headings assigned per record post 1979.

The three records with a very large number of DE terms (24 or 25) are given in Table 5, to illustrate some of LISA's indexing practices.

The first record, containing 25 nonduplicate terms, has an additional four superfluous DEs; the second record has two superfluous DEs; and the third record has none. There are some interesting features of the LISA chain indexing procedure (Edwards, 1975; Moore, 1988b) evident in the three records in Table 5. The chain indexing procedure is responsible for assigning both a DE term and either its Broader Term (BT) or Narrower Term (NT); for instance:

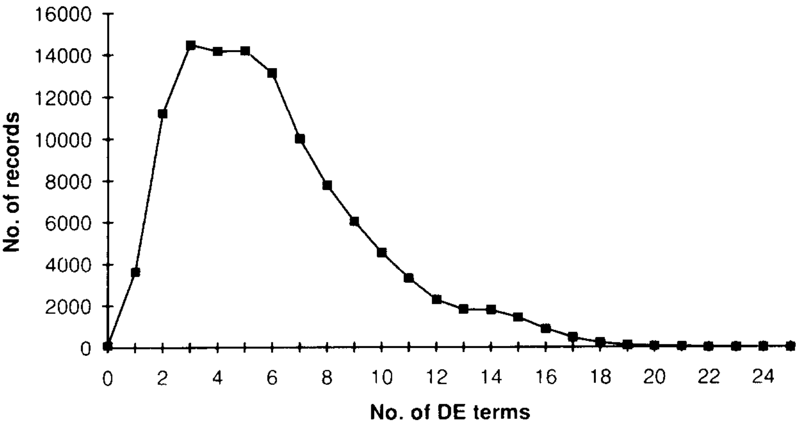

Fig. 5. Distribution of number of DE terms per record.

Table 5. LISA records with a high DE frequency

```
1.      Number of non-duplicate DEs = 25
TI: BIOSIS Previews & MEDLARS-a biomedical team
AU: Van-Camp,-Ann-J; Foreman,-Gertrude; Camp,-A-J-Van
SO: Online, 1 (1) Jan 77, 24-25, 27-29
DE: Minnesota-University; Indiana-University; U.S.A.-; Medical-libraries;
University-libraries; Multiple-data-base-searches; On-line-information-
retrieval; Searching-; External-magnetic-tape-information-services;
Information-services-Published-and-distributed-services; MEDLARS-; BIOSIS-
Previews; Medicine-; Biology-; Science-and-technology; Technology-and-
science; Magnetic-tape; Computerised-information-services; Technical-
processes-and-services; Information-storage-and-retrieval; Information-
retrieval; Subject-indexing; Computerised-information-retrieval;
Computerised-information-storage-and-retrieval; Searching-; Searching-;
Subject-indexing; Computerised-subject-indexing; Subject-indexing
DA: 1977

2.      Number of non-duplicate DEs = 24
TI: Conditions for loans to exhibitions
TO: Bedingungen fur Leihgaben zu Ausstellungen
AU: Kommission-fur-Handschriftenfragen],-[Verein-Deutscher-Bibliothekare;
Verein-Deutscher-Bibliothekare-(West-Germany)-Kommission-fur-
Handschriftenfragen; V-D-B-(Verein-Deutscher-Bibliothekare,-West-Germany)-
Kommission-fur-Handschriftenfragen; Kommission-fur-Handschriftenfragen-
Verein-Deutscher-Bibliothekare-(West-Germany); Commission-for-Manuscript-
Questions-Association-of-German-Librarians-(West-Germany); Association-of-
German-Librarians-(West-Germany)-Commission-for-Manuscript-Questions
SO: Z.-Biblioth.-u.-Bibliog., 24 (2) Mar-Apr 77, 96-104
DE: West-Germany; Regulations-; Antiquarian-materials; Old-and-Rare-
materials; Rare-and-Old-materials; Early-materials; Protection-; Security-;
Damage-; Hazards-; Preservation-; Conservation-; Interloans-; Cooperation-;
Displays-and-Exhibitions; Exhibitions-and-Displays; Library-publicity;
Publicity-Library-publicity; Kommission-fur-Handschriftenfragen-Commission-
for-Manuscript-Questions; Verein-Deutscher-Bibliothekare-Association-of-
German-Librarians-West-Germany; Services-; Reader-services; Use-promotion;
Publicity-; Displays-and-Exhibitions; Exhibitions-and-Displays
DA: 1977

3.      Number of non-duplicate DEs = 24
TI: The patron is not the public
AU: Hays,-Timothy; Shearer,-Kenneth-D; Wilson,-Concepcion
SO: Libr.-J., 102 (16) 15 Sept 77, 1813-1818. tables. 14 refs
DE: User-surveys; Surveys-; Non-users-; Community-information-services;
Neighbourhood-information-services; Welfare-services; North-Carolina;
Piedmont-area-North-Carolina; U.S.A.-; User-needs; Use-; Public-libraries;
Social-services; Postal-services; Mail-services; Books-by-mail-services;
Loans-; Telephones-; Reservation-systems; Requests-; Services-; Reader-
services; Information-work; Social-sciences
DA: 1978
```

Record 1. On-line-information-retrieval
Multiple-data-base-searches

Subject-indexing
Computerised-subject-indexing.

The two pairs of thesaurus terms above are generically reciprocal: In the first pair, "Multiple-data-base-searches" is the Narrower Term (NT), whereas in the second pair, "Subject-indexing" is the Broader Term (BT) (LISA, 1987, pp. 114, 201, 210, 259).

The first and second records have examples of word-order reversal; for example:

Record 1. Science-and-technology
Technology-and-science
Record 2. Displays-and-Exhibitions
Exhibitions-and-Displays.

In the first example, "Science-and-technology" is the preferred thesaurus term, while the reverse ("Technology-and-science") is a nonpreferred term (UF and USE reciprocal relationship). In the second example (unlike the first), both word orders are Thesaurus (preferred) terms (LISA, 1987, pp. 129, 140, 245, 264). These two terms in Record 2 account for the two superfluous DEs noted above.

The second record has in its DE field an example of an organization name with both the original-language and English-language equivalent given; for example:

Record 2. Kommission-fur-Handschriftenfragen-Commission-for-Manuscript-Questions

According to LISA's editor, Moore (1988b, p. 13), the removal of all names of systems and organizations from the Subject index (DE) would be effective from January 1979 onwards. Record 2 was entered into the LISA database in 1977.

In Record 3 the assignment of several general-to-specific place names in the DE field appears to be the indexing practice of LISA; for example:

Record 3. U.S.A.-
North-Carolina
Piedmont-area-North-Carolina.

Some of the indexing examples above depart from traditional indexing practice. Although some are the direct result of the chain indexing procedure, others appear to be the results of in-house indexing policies.

5.7 *First appearance (DA) of DE terms*

The date a DE term first appeared was analyzed. Because of the very high number of DEs used only once, it was decided to measure this feature at different levels of DE frequencies, to see if the frequently used terms followed the same trends as DEs overall. The thresholds of DE use chosen were: 1, 5, 10, 50, 100, 500, 1000, and 5000 terms. For example, in Table 6, the threshold of 500 represents all headings used 500 times or more in the

Table 6. Frequency distribution of first appearance of DE at varying thresholds

	DE term use threshold							
DA	**1**	**5**	**10**	**50**	**100**	**500**	**1000**	**5000**
1969	1764	729	539	285	212	104	59	7
1970	1159	192	108	40	28	10	7	1
1971	1208	130	70	34	19	4	2	0
1972	1495	112	65	20	9	1	0	0
1973	1515	89	64	24	14	4	0	0
1974	2418	534	371	164	115	27	15	2
1975	3095	729	471	155	77	17	6	1
1976	2431	390	232	87	63	27	15	8
1977	2626	419	235	43	21	4	1	0
1978	2198	173	73	12	6	1	1	0
1979	559	110	49	3	1	0	0	0
1980	834	131	65	14	8	3	0	0
1981	851	109	44	9	5	3	2	0
1982	873	91	33	6	2	1	1	0
1983	692	73	27	4	2	0	0	0
1984	578	47	20	4	3	1	1	0
1985	514	31	5	0	0	0	0	0
1986	600	32	11	4	3	2	2	0
1987	505	20	8	1	0	0	0	0
1988	489	19	7	1	1	0	0	0
1989	612	43	18	4	3	1	0	0
1990	694	8	0	0	0	0	0	0
1991	482	9	2	0	0	0	0	0
Total	**28192**	**4220**	**2517**	**914**	**592**	**210**	**112**	**19**

entire LISA database. In can be seen that in 1969, of all the 210 terms having a total frequency of over 500, nearly one half (104) of these headings were used for the first time in 1969.

Table 6 and Fig. 6 show that from 1974–1976, at any threshold level, a large number of new terms were added. For example, in 1976, eight of the nineteen terms (42.1%) occurring at least 5000 times first appeared in that year. Another way of interpreting the peaks during this period is that the indexing vocabulary used throughout the LISA database was largely established at this time. This was not expected. It was thought that the distribution would start with a large number of new terms as the database was established, and then decline with time to a low stable level. Instead, as can be seen, there were unexpected peaks in 1974–1976.

5.8 *DE term usage overlap*

The 17 most frequently used terms were analyzed for pairwise overlap in usage among all of them. In Table 7, the absolute overlap is given first, followed by the percentage overlap (i.e. the percentage of the overlap compared to the smaller of the two terms). For (2) INFORMATION-STORAGE-AND-RETRIEVAL, the absolute overlap with (1) TECHNICAL-PROCESSES-AND-SERVICES is 16926, for an overlap of 94%. The diagonals represent the frequency of occurrence of each of the top 17 headings.

Note the predominance of either a very high overlap (e.g., 100% between (2) INFORMATION-STORAGE-AND-RETRIEVAL and (3) INFORMATION-RETRIEVAL) or a very low, even non-existing overlap (e.g., 0% between (6) SUBJECT INDEXING and (5) SERVICES). The exceptions are (15) U.S.A. and (17) UNIVERSITY-LIBRARIES, which have a less skewed pairwise overlap distribution. (1) TECHNICAL-PROCESSES-AND-SERVICES occurs in 23% of all records (see Table 4) and in 17% of the records with the DE (15) U.S.A. This independence of the two terms can be contrasted with the opposite between pairs of terms such as (1) TECHNICAL-PROCESSES-AND-SERVICES and (2) INFORMATION-STORAGE-AND RETRIEVAL (94%), or between (2) INFORMATION-STORAGE-AND-RETRIEVAL and (3) INFORMATION-RETRIEVAL (100%). This dependence phenomenon can be largely explained by the chain indexing used in the LISA subject indexing procedure (Edwards, 1975; Moore, 1988b).

To explore LISA's chain indexing procedure further and LISA's theoretical rationale of subject indexing at all levels from specific to general or vice versa, the DE terms relating closely to the concept of information storage and retrieval, in its broadest sense, were analyzed. These include terms (2), (3), (6), (8), (12), (13), (14) in Table 7. It is apparent that in the range of overlap percentages (93%–100%), these DE terms are not only highly dependent on one another, but that they also form a hierarchy of terms, each of which can nearly substitute for the other. In looking at the LISA Thesaurus (LISA, 1987), the following interrelationships of the seven DE terms were noted.

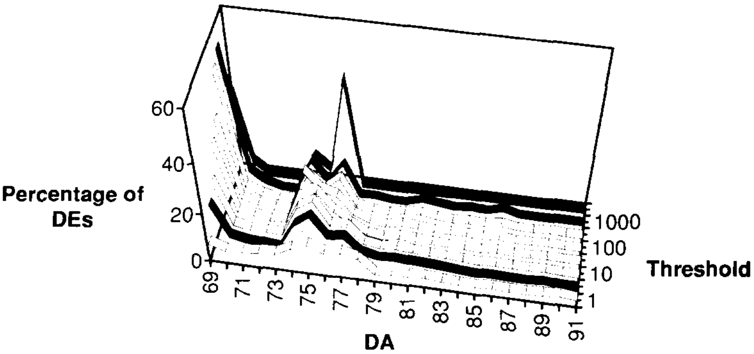

Fig. 6. Percentage of DEs by first appearance (DA) at varying thresholds.

Table 7. Usage overlap of 17 most frequently used DE terms

(1)	25942																
(2)	16926 94%	17922															
(3)	16926 94%	17917 100%	17919														
(4)	1013 6%	1 0%	1 0%	17361													
(5)	1011 6%	0 0%	0 0%	17347 100%	17347												
(6)	12275 91%	13262 98%	13262 98%	7 0%	7 0%	13512											
(7)	11 0%	10 0%	9 0%	12391 93%	12391 93%	10 0%	13313										
(8)	9205 91%	10147 100%	10147 100%	0 0%	0 0%	10139 100%	4 0%	10161									
(9)	896 9%	347 4%	347 4%	1114 11%	1112 11%	204 2%	428 4%	151 2%	9825								
(10)	311 3%	207 2%	207 2%	17 0%	16 0%	148 2%	11 0%	117 1%	1246 13%	9569							
(11)	310 3%	207 2%	207 2%	19 0%	18 0%	148 2%	11 0%	117 1%	1240 13%	9548 100%	9553						
(12)	8077 92%	8606 98%	8606 98%	19 0%	19 0%	8514 97%	9 0%	8144 93%	140 2%	120 1%	120 1%	8797					
(13)	7574 93%	8126 100%	8126 100%	0 0%	0 0%	8128 100%	2 0%	8126 100%	131 2%	114 1%	114 1%	8130 100%	8131				
(14)	7077 90%	7596 97%	7595 97%	74 1%	74 1%	7613 97%	25 0%	7572 97%	133 2%	134 2%	134 2%	7582 97%	7530 96%	7845			
(15)	1333 17%	738 10%	738 10%	1038 13%	1036 13%	434 6%	774 10%	356 5%	809 10%	662 9%	661 9%	331 4%	311 4%	318 4%	7765		
(16)	17 0%	7 0%	6 0%	113 2%	113 2%	6 0%	134 2%	4 0%	1427 19%	5 0%	4 0%	3 0%	3 0%	5 0%	661 9%	7364	
(17)	2458 36%	1165 17%	1165 17%	876 13%	876 13%	588 9%	495 7%	491 7%	37 1%	668 10%	666 10%	369 5%	332 5%	343 5%	490 7%	517 8%	6793
No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

Number	DE	Frequency
(1)	TECHNICAL-PROCESSES-AND-SERVICES	25942
(2)	INFORMATION-STORAGE-AND-RETRIEVAL	17922
(3)	INFORMATION-RETRIEVAL	17919
(4)	READER-SERVICES	17361
(5)	SERVICES	17347
(6)	SUBJECT-INDEXING	13512
(7)	INFORMATION-WORK	13313
(8)	COMPUTERISED-INFORMATION-STORAGE-AND-RETRIEVAL	10161
(9)	PUBLIC-LIBRARIES	9825
(10)	STOCK	9569
(11)	LIBRARY-MATERIALS	9553
(12)	SEARCHING	8797
(13)	COMPUTERISED-INFORMATION-RETRIEVAL	8131
(14)	ON-LINE-INFORMATION-RETRIEVAL	7845
(15)	U.S.A.	7765
(16)	ORGANISATION-AND-ADMINISTRATION	7364
(17)	UNIVERSITY-LIBRARIES	6793

The generic relationships of Broader Term (BT) and Narrower Term (NT) of three of the seven DE terms are as follows:

(2) INFORMATION-STORAGE-AND-RETRIEVAL
↓
(8) COMPUTERISED-INFORMATION-STORAGE-AND-RETRIEVAL
(14) ON-LINE-INFORMATION-RETRIEVAL.

Reciprocal BT/NT were given for terms (2) and (8). However, term (14) did not have term (8) as its BT; instead it had term (2) as a Related Term (RT). Likewise, term (2) had term (14) as an RT. The RT or associative relationship "is not a hierarchical relationship . . . it should *not* be used to link terms that appear in the same hierarchy" (Lancaster, 1986, p. 45). Additionally, the reciprocal BT/NT relationship should always be apparent (i.e., term (14) should have term (8) as its BT).

The two terms below:

(3) INFORMATION-RETRIEVAL
(13) COMPUTERISED-INFORMATION-RETRIEVAL

exist as "lone" terms without any generic or related terms. However, both are evidently a "part of" terms (2) and (8).

The remaining two terms:

(6) SUBJECT-INDEXING
(12) SEARCHING

are clearly a part of this subject cluster. In each case, the DE terms have as Used For (UF) the following:

(6) SUBJECT-INDEXING
UF INFORMATION-STORAGE-AND-RETRIEVAL by Subject Specification.
(12) SEARCHING
UF INFORMATION-RETRIEVAL: Searching
INFORMATION-STORAGE-AND-RETRIEVAL: Searching
RETRIEVAL:INFORMATION: Searching

The presence of the *words* INFORMATION RETRIEVAL or INFORMATION STORAGE AND RETRIEVAL somewhere in each thesaurus term entry suggests an overall relatedness, and perhaps even synonymy.

6. CLASSIFICATION CODES (CC)

The relationship between the Classification Code (CC) and the DEs assigned to a particular record was examined. Because the chain indexing used by LISA to produce the DE terms is based on the Classification Research Group (CRG) faceted scheme, it was felt that a mapping of an assigned CC to the DEs in the same record would give a fair understanding of how the DEs are generated from the chain indexing procedure (Edwards, 1975, pp. 135–140). Because the chain indexing procedure used by LISA originates from the CC assigned to each record, it was important to examine at least some of the CCs and their corresponding (semi-automatically generated) DEs. Table 8 looks at the 22 of a total of 59 records (DA for the first three years 1981–1983 and the last, 1991), which were assigned the CC of **ZmNx Zm**. The "No." column represents the number of exact mappings of the CC to the DEs assigned in any given DA year. Some combinations of headings appear uniquely (e.g., in 1982 beginning with "Satellites"), whereas other combinations appear more than once (as in the first example, which appears three times). Although it is apparent that there is a strong relationship between the CC and the set of DEs assigned to each of the 22 records, there is not always a total correlation. Of particular interest is the fact that records assigned this CC in 1991 no longer have the very broad DE term **Technical-processes-and-services**. It can also be seen that some of the variety of DEs appearing in only some and not all of the 22 records are specific place names (e.g., Brazil); although there are others that are subject terms (e.g., Local-area-networks and Satellites). In at least two DAs (1981, 1983), the mapping of the CC to its corresponding DE terms was totally consistent, even to the duplicated DEs (e.g., "Searching"). In the remaining DA years, however, the variations of mapping procedure could not be systematically determined. It is

Table 8. Twenty-two of the fifty-nine records with the CC "ZmNxZm" and the corresponding DEs

CC	DA	No.	Corresponding DE terms assigned
ZmNx Zm	1981	3	Technical-processes-and-services; Information-storage-and-retrieval; Information-retrieval; Subject-indexing; On-line-information-retrieval; Computerised-information-retrieval; Searching-; Searching-; Computerised-information-storage-and-retrieval; Telecommunications-; Data-transmission
ZmNx Zm	1982	1	Satellites-; Technical-processes-and-services; Information-storage-and-retrieval; Information-retrieval; Subject-indexing; On-line-information-retrieval; Computerised-information-retrieval; Searching-; Searching-; Computerised-information-storage-and-retrieval; Telecommunications-; Data-transmission
ZmNx Zm	1982	2	Technical-processes-and-services; Information-storage-and-retrieval; Information-retrieval; Subject-indexing; On-line-information-retrieval; Computerised-information-retrieval; Searching-; Searching-; Computerised-information-storage-and-retrieval; Telecommunications-; Data-transmission
ZmNx Zm	1983	7	Technical-processes-and-services; Information-storage-and-retrieval; Information-retrieval; Subject-indexing; On-line-information-retrieval; Computerised-information-retrieval; Searching-; Searching-; Computerised-information-storage-and-retrieval; Telecommunications-; Communications-; Data-transmission
ZmNx Zm	1991	2	Information-storage-and-retrieval; Information-retrieval; Subject-indexing; On-line-information-retrieval; Computerised-information-retrieval; Searching; Searching; Computerised-information-storage-and-retrieval; Telecommunications; Data-transmission
ZmNx Zm	1991	1	Information-storage-and-retrieval; Information-retrieval; Subject-indexing; On-line-information-retrieval; Computerised-information-retrieval; Searching; Searching; Computerised-information-storage-and-retrieval; Telecommunications; Data-transmission; Brazil
ZmNx Zm	1991	1	Information-storage-and-retrieval; Information-retrieval; Subject-indexing; On-line-information-retrieval; Computerised-information-retrieval; Searching; Searching; Computerised-information-storage-and-retrieval; Telecommunications; Data-transmission; Hungary
ZmNx Zm	1991	1	Information-storage-and-retrieval; Information-retrieval; Subject-indexing; On-line-information-retrieval; Computerised-information-retrieval; Searching; Searching; Computerised-information-storage-and-retrieval; Telecommunications; Data-transmission; Planning; Local-area-networks
ZmNx Zm	1991	1	Information-storage-and-retrieval; Information-retrieval; Subject-indexing; On-line-information-retrieval; Computerised-information-retrieval; Searching; Searching; Computerised-information-storage-and-retrieval; Telecommunications; Data-transmission; Satellites
ZmNx Zm	1991	1	Information-storage-and-retrieval; Information-retrieval; Subject-indexing; On-line-information-retrieval; Computerised-information-retrieval; Searching; Searching; Computerised-information-storage-and-retrieval; Telecommunications; Data-transmission; U.S.S.R.; Design
ZmNx Zm	1991	1	Information-storage-and-retrieval; Information-retrieval; Subject-indexing; On-line-information-retrieval; Computerised-information-retrieval; Searching; Searching; Computerised-information-storage-and-retrieval; Telecommunications; Data-transmission; User-System-interface; Electronic-mail; Mail
ZmNx Zm	1991	1	Information-storage-and-retrieval; Information-retrieval; Subject-indexing; On-line-information-retrieval; Computerised-information-retrieval; Searching; Searching; Computerised-information-storage-and-retrieval; Telecommunications; Data-transmission

unclear exactly what chain indexing procedure was used by the LISA database through the years. However, one thing that is clear to us is that in any one DA year, the same CC can generate a set of DEs in one record that is not shared by other records with the same DA or entered in the database at an earlier or later time!

In trying to gather information on the indexing procedure and policy used in the LISA database, we wrote to the present editor-in-chief, N. L. Moore. In response to our request for some information, we were indirectly told that what is in the public literature we had already obtained (see References) was all that could be provided. Mr. Moore did go on to say, however, that:

> A glimmer of hope for the future lies in the fact that a new thesaurus will be compiled this year [1992] but it will be for indexing current issues and will not include the kind of historical notes about changes to indexing that the *LISA Online User Manual* did. Lastly . . . the new indexing system is . . . going to be a chain indexing system and is going to be similar to that used in *Current Technology Index* only leading to abstract numbers.

In closing Mr. Moore indicated that perhaps the changes that will appear in LISA from January 1993 onward could form an interesting research project (N.L. Moore, personal communication, April 21, 1992).

7. CONCLUSION

Our analysis of the assigned subject fields (CCs and DEs) in the LISA database has led to the conclusion that LISA's indexing policy and practice is far from exemplary. The thesaurus (or controlled vocabulary) LISA (1987) uses can quickly be seen as one not compiled according to standard thesaurus construction principles. The first edition of the *LISA Online User Manual* (1982), which calls the "thesaurus" a "Preferred Terms List," is a more honest assessment of the source of DE terms in the LISA database. We certainly look forward to the new thesaurus and chain indexing system to be adopted in the indexing of new records entered in the database from 1993 onwards. It can only be an improvement in what has been used until now.

The next phase of our study on LISA will focus on further analysis of the DE and CC fields with respect to term dependencies and term clustering over time. This type of investigation could reveal interesting results, such as a description of the evolution of the field of Library and Information Science, as well as the effects on retrieval performance of what appears to be an unusual pattern of term dependencies. We also intend to apply similar analysis to other databases that use controlled vocabulary in their indexing to refine our methods as a tool for comparison of the subject indexing in various databases.

Acknowledgements—The authors would like to thank the anonymous referees for their useful comments, which helped in the revision of this paper.

REFERENCES

Bottle, R.T., & Efthimiadis, E.N. (1984). Library and information science literature: Authorship and growth patterns. *Journal of Information Science*, *9*(3), 107-116.

Chu, C.M., & Ajiferuke, I. (1989). Quality of indexing in library and information science databases. *Online Review*, *13*(1), 11-35.

Day, J.M. (1988). LISA on CD-ROM—A user evaluation. In *Online information 87* (pp. 273-284). Oxford, U.K.: Learned Information.

Edwards, T. (1975). Indexing LISA: Chains, KISS and the bold approach. *The Indexer*, *9*(4), 133-146.

Edwards, T. (1976). LISA: A traditional abstracting service? *International Forum for Information and Documentation*, *1*(2), 25-34.

Feng, S. (1989). A comparative study of indexing languages in single and multidatabase searching. *Canadian Journal of Information Science*, *14*(2), 26-46.

Gini, C. (1909). Il diverso accrescimento delle classi sociali e la concentrazione della ricchezza. *Giornale degli Economisti*, Serie 11, 37.

Hartley, R.J. (1989). LISA on CD-ROM: An evaluation. *Online Review*, *13*(1), 53-56.

Hood, W., & Wilson, C.S. (1992). *An analysis of the indexing used in the LISA database.* Kensington, Australia: The School of Information, Library and Archive Studies, University of New South Wales.

Lancaster, F.W. (1972). *Vocabulary control for information retrieval.* Washington, D.C.: Information Resources Press.

Lancaster, F.W. (1986). *Vocabulary control for information retrieval*, 2nd edition. Washington, D.C.: Information Resources Press.

LISA, 1969-September 1991 CD-ROM edition. (1991). Boston, MA: SilverPlatter.

LISA Online User Manual (1982). Oxford, U.K.: Learned Information.

LISA Online User Manual, 2nd edition. (1987). Oxford, U.K.: Learned Information.

Moore, N.L. (1988a). Searching LISA on the SilverPlatter CD-ROM system. *Program*, *22*(1), 72-76.

Moore, N.L. (1988b). LISA indexing: Economic aspects of controlled indexing. *The Indexer*, *16*(1), 11-16.

Nelson, M.J., & Tague, J.M. (1985). Split size-rank models for the distribution of index terms. *Journal of the American Society for Information Science*, *36*(5), 283-296.

Pratt, A.D. (1977). A measure of class concentration in bibliometrics. *Journal of the American Society for Information Science*, *28*(5), 285-292.

Sievert, M.E., & Verbeck, A. (1987). The indexing of the literature of online searching: A comparison of ERIC and LISA. *Online Review*, *11*(2), 95-104.

Terbille, C. (1990). LISA on CD-ROM. *CD-ROM Librarian*, *5*(8), 26-28.

Tomlinson, D.M. (1986). LISA: anatomy of an abstracting service. *The Indexer*, *15*(2), 83-86.

APPENDIX-LISA ADJUSTED DE COUNTS

Year added (DA)	**Records per DA**	**Total adjusted DE count** (inc. "no heading")	**No. records with "no heading"**	**Actual adjusted DE count**	**Ave. adjusted DE per record**	**DEs used once**		**DEs used twice**	
						No.	**%**	**No.**	**%**
1969	2513	5540	0	5540	2.2	682	12.3	228	4.1
1970	2680	6073	0	6073	2.3	703	11.6	250	4.1
1971	2480	5773	0	5773	2.3	823	14.3	280	4.9
1972	2979	7148	0	7148	2.4	1125	15.7	361	5.1
1973	2853	7210	0	7210	2.5	1240	17.2	400	5.6
1974	2384	10351	5	10346	4.3	1474	14.3	396	3.8
1975	3848	16793	81	16712	4.3	1662	9.9	666	4.0
1976	3781	28516	1	28515	7.5	1477	5.2	652	2.3
1977	3917	32140	1	32139	8.2	1516	4.7	686	2.1
1978	3853	32160	2	32158	8.4	1627	5.1	672	2.1
1979	3813	23938	0	23938	6.3	296	1.2	131	0.6
1980	4887	31150	1	31149	6.4	523	1.7	182	0.6
1981	6020	38688	5	38683	6.4	505	1.3	209	0.5
1982	6004	36562	2	36560	6.1	581	1.6	208	0.6
1983	6778	47725	1	47724	7.0	422	0.9	222	0.5
1984	6992	44696	2	44694	6.4	395	0.9	176	0.4
1985	6506	41447	2	41445	6.4	375	0.9	178	0.4
1986	6475	40382	1	40381	6.2	442	1.1	225	0.6
1987	6433	40106	2	40104	6.2	384	1.0	175	0.4
1988	6498	42283	1	42282	6.5	381	0.9	192	0.5
1989	6488	42436	0	42436	6.5	440	1.0	180	0.4
1990	8141	54114	0	54114	6.7	580	1.1	292	0.5
1991	5130	34280	1	34279	6.7	434	1.3	217	0.6
Total	**111453**	**669511**	**108**	**669403**	**6.0**	**18087**	**2.7**	**7178**	**1.1**