Testing a New Design for Subject Access to Online Catalogs

by
Karen M. Drabenstott and Marjorie S. Weller
with the research assistance of Jeffrey M. Holden

School of Information and Library Studies
The University of Michigan
304 West Engineering Building
550 East University Avenue
Ann Arbor, Michigan 48109–1092 USA

January 1995

Testing a New Design for Subject Access to Online Catalogs

by
Karen M. Drabenstott and Marjorie S. Weller
with the research assistance of Jeffrey M. Holden

School of Information and Library Studies
The University of Michigan
304 West Engineering Building
550 East University Avenue
Ann Arbor, Michigan 48109–1092 USA

January 1995

Karen M. Drabenstott and Marjorie S. Weller

Table of Contents

List of Tables

List of Figures

Acknowledgments

The ASTUTE project team consisted of Karen Markey Drabenstott, Marjorie S. Weller, and Jeffrey M. Holden.

In this report, we describe a test of a new design for subject access to online catalogs. The Department of Education provided support for the test. We are especially grateful to Neal K. Kaske, Senior Associate, Office of Library Programs, Department of Education, who encouraged our research efforts and answered questions that arose during the project.

There were many staff members at participating libraries who contributed to the success of this Department of Education-sponsored research project. Timothy F. Richards, Head, Mardigian Library, University of Michigan-Dearborn (UM-D), and Evan Ira Farber, Director, Lilly Library, Earlham College, welcomed us to their libraries to collect bibliographic data and conduct search experiments with their libraries' patrons and staff. Robert Kelly at UM-D and Michael Bowden at Earlham helped us install the experimental online catalog, monitored the system during its installation at their libraries, transferred or sent transaction logs to us electronically or via ground mail, recruited library staff to participate in the search experiment, and answered our many questions about their libraries' bibliographic records and users. When we realized our first data collection effort at UM-D had been unsuccessful, both Tim and Bob invited us back to Mardigian Library to conduct a second search experiment. When we experienced difficulty in the speed of downloading bibliographic records from Earlham's online catalog database to floppy disk, Michael loaned us his new, speedy microcomputer for several days for downloading tasks. We also thank library staff members at UM-D and Earlham for taking part in the search experiment.

Advice from our project consultants was encouraging and supportive throughout the project. We appreciate the efforts of Martin Dillon and Paul Kantor who advised us on the experimental design. Paul Kantor reviewed specifications for the transaction logging capability and made suggestions on analyzing logged data. William H. Mischo

and Paul Kantor reviewed system functionality, drafts of the final report, and offered helpful suggestions on each.

At the Library of Congress, Susan Tarr of Cataloging Distribution Service made it possible for us to obtain a complimentary copy of CD/MARC Subjects which we put to work in the experimental online catalog.

At the University of Michigan-Ann Arbor, we appreciate the assistance of Michael G. Moore, presently Systems Development Coordinator, Population Studies Center, who answered technical questions about FoxPro, helped us download database tapes from UM-D, and transferred the ASTUTE development application into a user-executable application. We thank Jeffrey M. Holden who submitted transaction log data to statistical tests for the statistical and failure analyses. The authors are also grateful to Robert Royce who transferred and converted screen images from FoxPro to Microsoft Word and provided expert assistance in the design and layout of this report.

Obtaining Copies of this Report

[Karen — fill in at end.]

Karen M. Drabenstott and Marjorie S. Weller

About the Authors

Karen Markey Drabenstott is an Associate Professor in the School of Information Studies at the University of Michigan. The impetus for the research discussed in this study were findings from a Council on Library Resources-sponsored research project which are given in the book entitled *Using Subject Headings for Online Retrieval: Theory, Practice, and Potential* written by Karen Drabenstott and Diane Vizine-Goetz and published by Academic Press in 1994. Support from the Council also enabled Karen to research and write the *Analytical review of the library of the future* in 1994. This analytical bibliography and synthesis of published literature on the library of the future prepared her for her current role as faculty coordinator of the Kellogg Coalition on Information Science, Technology, and Library Education (KRISTaL-Ed), a five-year, multi-million dollar project supported by the Kellogg Foundation to provide national leadership in educating information professionals for the 21st century.

Karen joined the faculty of The University of Michigan in January 1987. From 1981 to 1986, she was a research scientist in the Office of Research at OCLC. She received her B.A. from The Johns Hopkins University and her M.L.S. and Ph.D. from the School of Information Studies at Syracuse University.

Marjorie S. Weller designed, developed, and implemented the ASTUTE experimental online catalog using the FoxPro database management system. She is currently a Programmer Analyst I in the Medical Center Information Technologies where she programs financial information systems for the Medical School of the University of Michigan. She received an A.S. in Computer Science from Henry Ford Community College.

Jeffrey M. Holden generated the statistics for the quantitative analysis of online retrieval test data. He is currently [Karen-fill in later]. He holds an A.B. and M.B.A. from Michigan State University and is pursuing a doctorate in the School of Information and Library Studies.

Karen M. Drabenstott and Marjorie S. Weller

1 Project Objectives and Research Questions

The 1980s began with a handful of libraries offering end users direct computerized access to library holdings through the online catalog component of their local online system. Ensuing the introduction of online catalogs in libraries were studies of online catalog users. One of the earliest and most comprehensive studies was a nationwide survey of online catalog use in twenty-nine libraries sponsored by the Council on Library Resources (CLR) (Matthews et al. 1983). The remark "Sacred cows are being strewn all over the landscape" characterized survey findings because they challenged traditional beliefs about catalog users and uses (Besant 1982, 160). In particular, survey findings about subject access — the predominance of subject searching, the many users who experienced difficulties conducting subject searches, and the large number of suggestions from users for improvements to subject searching — were completely unexpected.

Nationwide survey findings stimulated a decade of research in the area of subject access. CLR, OCLC (Online Computer Library Center), BLRDD (British Library Research and Development Department), and the (U.S.) Department of Education have been exemplary in their support of research on subject access (Drabenstott 1991, 64–6). Library staff also called for improvements to online catalogs to reflect user needs. Since staff play a large role in system selection, system designers have been especially receptive to staff demands for improved ease of use, enhancements to search functionality (i.e., keyword, Boolean-based approaches), and extensions of online catalog functionality to other databases (Hildreth 1991, 21).

Online catalogs are now commonplace in academic and public libraries. Telecommunications make it possible for end users to access a library's holdings without having to set foot in the library. National and international computer networks enable users to access online catalogs of libraries throughout the United States and the world.

Despite a decade of online catalog research, development, and deployment, many of the same problems that the earliest online catalog searchers experienced plague today's users, particularly in the area of subject access. Examples are:

- One-third of the subject queries users enter into online catalogs fail to produce retrievals. (Lynch 1989, 52)
- When searches produce retrievals, large retrievals discourage users from scanning results. (Markey and Demeyer 1986, 277; Van Pulis and Ludy 1988, 528; Lynch 1989, 52)
- The few instances when users are successful matching the catalog's controlled vocabulary are when they enter one-word queries for topics or places. (Van Pulis and Ludy 1988, 527; Carlyle 1989, 44; Drabenstott and Vizine-Goetz, 1994, 168)
- Users have become so discouraged with the results of subject searches that they are seeking alternative approaches to those that manipulate the subject headings in cataloging records. (Larson 1991)

These problems are indicative of the need for a new design for subject access to online catalogs. The foundation for the new design are findings from an empirical study of the subject terms users enter into online catalogs (Drabenstott and Vizine-Goetz 1994). The new design requires that online catalogs have a wide range of subject searching capabilities and search trees to govern the system's selection of searching capabilities in response to user queries. Search trees hold much promise for assuming the burden of determining which subject searching approach is likely to produce useful information in response to user queries.

This report describes a research project that tested the new subject access design. Project objectives were to:

1. Demonstrate subject searching in an experimental online catalog with a wide range of subject searching capabilities and with search trees to govern system selection of a particular capability in response to user queries.
2. Test the retrieval effectiveness of the experimental online catalog with search trees by comparing its performance with the performance of an experimental online catalog in which subject searching approaches are assigned at random.
3. Evaluate the demonstration and test results of retrieval effectiveness and disseminate the research findings through publications in the professional literature.

Underlying the design of the ASTUTE experimental online catalog were two important assumptions: (1) subject searching functionality was limited to searching capabilities in operational online catalogs, and (2) titles and subject headings (based on

the Library of Congress Subject Heading system) were the principal indicators of subject content in machine-readable cataloging (MARC) records.

The results of the research project have the potential for generating a new design. Since ASTUTE's subject searching functionality was limited to capabilities in operational catalogs, system designers would not have to discard their operational online catalogs because the new design would be based on subject searching approaches that are already available in existing systems. Similarly, the subject content of ASTUTE's cataloging databases was based on the same information in the cataloging databases of most American libraries. Thus, library technical services staff would not have to discard their existing databases or introduce new methods to take advantage of the new design. The new design would, however, require system designers to add a wider range of subject searching functionality than their systems currently support and search trees to control system responses.

The research project was composed of twelve activities grouped into three phases: (1) system development (tasks 1–6, 9), (2) system testing (tasks 7, 8, 10), and (3) data analysis and reporting (tasks 11–12). Figure 1.1 shows a flowchart of project tasks and gives a schedule of project tasks.

Seven activities (1–6, 9) were required in the system-development phase. This phase featured the development of an experimental online catalog with a wide range of subject searching functionality and search trees to govern the system's selection of a subject searching approach in response to user queries. Bibliographic records from the online catalogs of Mardigian Library at the University of Michigan-Dearborn (UM-D) and Lilly Library at Earlham College were loaded into separate, searchable databases of the experimental online catalog. Three activities (7–8, 10) were required in the system testing phase. Data collection procedures and instruments were developed, pretested, revised, and administered in online retrieval experiments with library patrons and staff at the libraries of UM-D and Earlham College to answer the following three research questions:

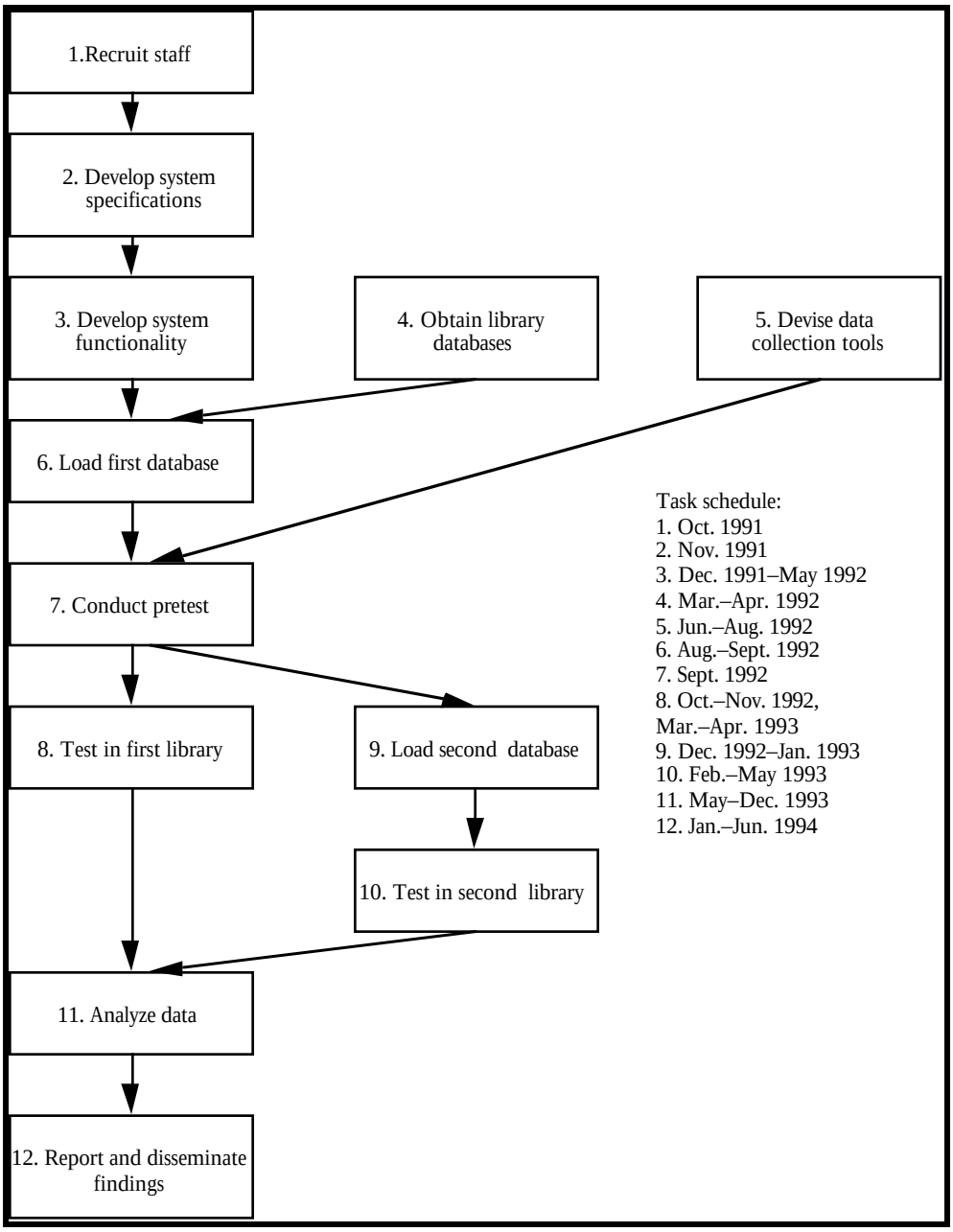

Figure 1.1. Project activities

1. Do search trees improve the search performance of subject searchers at the online catalog?
2. Do subject searchers prefer an online catalog that controls system responses and searching approaches to an online catalog without such controls?

3. What are the characteristics of queries that cannot be answered using the ASTUTE experimental online catalog?

Two activities (11–12) made up the evaluation phase of the project. Collected data were described, analyzed, interpreted, and synthesized in a final report to the Department of Education. The ASTUTE Project Team also plans to disseminate findings in journals and conference proceedings to the library and information science communities.

This report of the ASTUTE Project consists of thirteen chapters. Chapter 1 describes factors that led to the project director's research proposal on testing a new subject access design to online catalogs and the research questions and objectives of the project. Chapter 2 reviews previous research on search trees, subject searching functionality, and user queries for subjects. Chapters 3–5 describe the development of the experimental online catalog, in particular, its bibliographic and authority databases, subject searching functionality, and search trees. Chapter 6 focuses on the methods used in online retrieval tests and includes a discussion of the catalog's transaction logging capability that was used to record users' subject searches, relevance assessments, and administer pre- and post-search questionnaires. The results of the analysis of online retrieval test data are covered in chapters 7–12. Results chapters include recommendations on reconfiguring and enhancing search trees to accommodate the wide range of subject queries users enter into online catalogs. Chapter 13 is a summary of project activities, findings, and recommendations.

The data and analyses of this research project generally increase our understanding of the subject terms users enter into online catalogs. The project builds on a new design for subject access to online catalogs that enlists search trees to control system responses and determine appropriate subject searching approaches to user queries. The project will ensure that future subject searching implementations in online catalogs are responsive to the wide variety of user queries for subjects.

References

Besant, Larry. 1982. "Early survey findings: users of public online catalogs want sophisticated subject access." *American Libraries* 107, 19 (March): 160.

Carlyle, Allyson. 1989. "Matching LCSH and user vocabulary in the library catalog." *Cataloging & Classification Quarterly* 10, (1/2): 37–63.

Drabenstott, Karen M. 1991. "Online Catalog User Needs and Behaviors." In *Think Tank on the Present and Future of the Online Catalog: Proceedings*, edited by Noelle Van Pulis, 59–83. Chicago: Reference and Adult Services Division, American Library Association. RASD Occasional Papers, 9.

Drabenstott, Karen M., and Diane Vizine-Goetz. 1994. *Using subject headings for online retrieval: Theory, practice, and potential*. San Diego: Academic Press.

Drabenstott, Karen M., and Diane Vizine-Goetz. 1990. "Search trees for subject searching in online catalogs." *Library Hi Tech* 8, 3: 7–20.

Hildreth, Charles R. 1991. "Advancing toward the E[3] OPAC: the imperative and the path." In *Think Tank on the Present and Future of the Online Catalog: Proceedings*, edited by Noelle Van Pulis, 17–38. Chicago: Reference and Adult Services Division, American Library Association. RASD Occasional Papers, 9.

Larson, Ray R. 1991. "The Decline of Subject Searching: Long-term Trends and Patterns of Index Use in an Online Catalog." *Journal of the American Society for Information Science* 42, 3 (April): 197–215.

Lynch, Clifford A. 1989. "Large Database and Multiple Database Problems in Online Catalogs." In *OPACs and beyond*. Dublin, Ohio: OCLC.

Markey, Karen, and Anh Demeyer. 1986. *Dewey Decimal Classification Online Project; Final Report to the Council on Library Resources*. Dublin, Ohio: OCLC. OCLC Research Report OCLC/OPR/RR–86/1.

Matthews, Joseph R., et al. 1983. *Using Online Catalogs: A Nationwide survey*. New York: Neal-Schuman.

Van Pulis, Noelle, and Lorene Ludy. 1988. "Subject Searching in an Online Catalog with Authority Control." *College & Research Libraries* 49, 6 (November): 523–33.

2 Literature Review

2.1 Studies of the Subject Terms Users Enter into Online Catalogs

The computer's ability to record every system response and user action input into the online catalog provides researchers with a very accurate tool for collecting the subject terms users enter into online catalogs. Such user-system interaction is called a transaction log.

At the present time, standards do not exist for transaction logs. Consequently, the content and format of transaction logs varies from system to system. Computer programs written to analyze transaction log data from one system must be rewritten to analyze data from a different system.

An important advantage of transaction logs is the unobtrusiveness of this data collection approach. A computer program collects user-system interaction data unbeknownst to catalog users. Consequently, users are not likely to alter their catalog-seeking behavior because they do not know that the terms they are entering into the catalog are being recorded for analysis. Another advantage of transaction logging is the ability to collect voluminous amounts of data in machine-readable form for subsequent analyses using computer programs.

Transaction logs also have disadvantages. Logged queries do not always provide researchers with an exact representation of what users are looking for. This information could be obtained by supplementing transaction logs with observations of user searches, personal interviews, or a combination of the two. Logs also do not provide researchers with accurate demarcations of searches entered by different users. Researchers have undertaken manual analyses of transaction logs and considered changes in both time stamps and the meaning of user queries to demarcate individual searches. Hildreth (1985) and Kurth (1994) commented on these and other disadvantages of transaction logging. To overcome disadvantages of this data collection method, researchers in the United Kingdom advocated the utilization of a frontend

system with four capabilities: (1) full-screen logging, (2) playback facility, (3) pre- or post-search, online or offline questionnaire administration, or (4) in-search questionnaire administration (Hancock et al. 1990).

This chapter summarizes what has been learned about the subject terms users enter into online catalogs. Findings are given in the form of generalizations about the subject terms users enter into online catalogs. Studies were limited to those in which researchers selected a random sample of user queries from transaction logs or selected user queries from a transaction log recorded over a certain time period. Studies in which researchers gave respondents a search task to perform and recorded their activity on a transaction log were not included because of the obtrusiveness of this data collection approach.

2.1.1 What Have We Learned about User Queries for Subjects?

In *Using subject headings for online retrieval* (1994), Karen Drabenstott and Diane Vizine-Goetz summarized what had been learned to date about user queries for subjects from online catalog use studies in the form of twenty generalizations. This section gives brief discussions of generalizations. It cites studies that support or fail to support generalizations in the form of study numbers and page numbers; the former correspond to the studies listed in the "References" section that concludes this chapter. (The reference lists of studies that support or fail to support generalizations include a few more studies that have been published since the completion of *Using subject headings for online retrieval*.)

Generalization	Studies
1. A small proportion of the terms users enter into online catalogs are not legitimate subject queries.	Supporting: 16:66; 8:48; 23:76; 3:51; 15:172; 19:270; 12:400; 24:53; 1:68; 5:160–3.

Queries that were not legitimate subject queries were random configurations (///,HJNVM), data entry errors, and expletives. Such activity could be exploratory, accidental, or indicative of the frustration users are experiencing with their online search. Researchers have used different terminology to describe these terms, e.g., malicious entries, garbage, entry errors, graffiti, identifiable obscenities, questions, and comments. The important point about this activity was that it occurred with much less frequency than the entry of legitimate subject queries. For example, such activity was only 0.4% of the random sample of subject queries extracted from transaction logs at Northwestern University (Lester 1989, 172). Thus, online catalog searchers understood enough about the system to enter a legitimate subject search term, i.e., a term that had the potential to retrieve bibliographic records.

Generalization	Studies
2. Users include punctuation in their queries.	Supporting: 8:50; 23:78; 15:184.

Examples of such punctuation were possessive forms with an apostrophe, acronyms with periods in between letters, hyphenated words and phrases, and inverted phrases entered with an intervening comma.

Generalization	Studies
3. Users enter phrases without intervening spaces.	Supporting: 8:48; 13:6; 23:76; 1:68; 5:176.
4. Users misspell words.	Supporting: 16:66; 8:48; 13:4; 23:76; 15:194, 197; 19:170; 12:400; 24:53; 11:28; 2:41; 1:68; 5:175–6; Failing to support: 3: 44.

Spelling errors — substituting, inserting, transposing, and omitting one or more letters — prevented user queries from matching words and phrases indexed in online catalogs. Although spelling errors occur, they were not very prevalent in user queries. For example, about 5% of the subject queries examined in separate studies by Lester (1989, 194, 197) and Jones (1986, 4) contained spelling errors.

Generalization	Studies
5. Users enter abbreviations.	Supporting: 8:48; 23:76; 3:44.

Abbreviated terms were not that common; in fact, Carlyle (1989, 44) only found one such term in her sample of 161 subject terms.

Generalization	Studies
6. Users preface their queries with an initial article.	Supporting: 12:400; 2:41; 1:68.

Initial articles can prevent systems from retrieving bibliographic records unless systems are programmed to delete initial articles from user queries.

Generalization	Studies
7. Users enter queries in singular form when the authorized form in LCSH is plural and vice versa.	Supporting: 3:44; 2:44; 5:177–80.

Barrett and Maticka (1990, 44) gave examples of singular subject headings (“Kimono” and “Wheel”) for which users entered plural forms and of plural subject headings (“Apples” and “Kidney beans”) for which users entered singular forms.

Generalization	Studies
8. Users enter queries containing a date that differs in form from the date in the subject heading in the LCSH system.	Supporting: 3:44.

An example was the user query containing the date "20th century." (LCSH represents this date as "Twentieth century.")

Generalization	Studies
9. Users enter queries containing a different suffix than the subject heading in the LCSH system.	Supporting: 3:44.

An example was the user query containing the suffix "computer programming" that was different from the suffix of the subject heading based on LCSH, viz. "Computer programs."

Generalization	Studies
10. Users enter queries using the online catalog's subject search capability that are probably better suited to the catalog's author or title capability.	Supporting: 19:270; 12:400; 24:53; 5:159–61.

Users who entered such queries could have been unsure which searching capability they should use to enter their query (so they used all of them) or they could havebeen surveying results to determine which capability provided a manageable number of useful retrievals. Fortunately, such queries were not very common.

Generalization	Studies
11. Users enter queries subject queries that match *see* references in the Library of Congress Subject Headings.	Supporting: 16:66; 15:239; 2:41; 5:169.

Lester (1989, 239) reported that 9.8% of subject queries entered into the NOTIS online catalog at Northwestern University matched *see* references in LCSH. Corrections to spelling errors resulted in an additional 0.8% of matches.

Generalization	Studies
12. Users enter queries subject queries that exactly match authorized headings in the Library of Congress Subject Headings (LCSH).	Supporting: 15:172; 6:108; 2:41; 5:167–70, 187–8.

Lester (1989, 172) reported that 40% of the subject queries entered by searchers of the NOTIS online catalog at Northwestern University were exact matches of the subject

headings printed in *LCSH*. (She disregarded capitalization and punctuation, i.e., commas, periods, dashes, hyphens, apostrophes, semicolons, and parentheses, in the matching process.)

Generalization	Studies
13. Users enter subject queries that exactly match subject headings in a bibliographic database based on the LCSH system.	Supporting: 16:66; 3:44; 19:269; 12:399; 24:53; 2:41.

Percentages of matching queries ranged from a low of 18% (Markey 1984, 66) to high of 48% (Peters 1989, 269). The different methods used in the five studies to arrive at percentages of matched queries could explain the wide disparity in percentages.

Generalization	Studies
14. Users that are exact matches of the controlled vocabulary are likely to be shorter than nonmatches.	Supporting: 16:67, 70. Failing to support: 15:176.

Lester (1989, 176) reported that exact matches in her study were the same length (1.7 words) as nonmatches. The nonmatches in Markey's (1984, 67, 70) study were longer (2.3 words) than exact matches (1.5 words).

Generalization	Studies
15. The majority of the subject queries users enter into online catalog's using the system's subject searching capabilities are for topical subjects.	Supporting: 15:186; 5:158–9.

Over 70% of the queries in Lester's (1989, 186) study were for topical subjects. A surprising high percentage (44%) of these queries matched subject headings printed in *LCSH*. The percentage (62.2%) reported by Drabenstott and Vizine-Goetz was comparable to the percentage reported by Lester.

Generalization	Studies
16. Queries for names or combinations of names and other words are less likely to match subject headings printed in *LCSH* than queries for topical subjects.	Supporting: 15:188; 5:225–9.

Of the nearly 29% of user queries in Lester's (1989, 188) study that contained personal, corporate, or geographic names, only thirty percent were successful in matching a subject heading printed in *LCSH*. The high percentage of nonmatches could be explained by the exclusion from LCSH of most headings for personal and

corporate names and geographic places that name jurisdictional units. (Such headings are more likely to found in the Library of Congress Name Authority File (LCNAF) than in LCSH.)

Generalization	Studies
17. User queries are sometimes an amalgamation of two or more subject headings from the LCSH system.	Supporting: 3:44.

An example was the user query "crystallography geometry" that matched two separate subject headings printed in *LCSH*: "Crystallography" and "Geometry."

Generalization	Studies
18. The retrieval processes in online catalogs can increase the match success rate between user-entered terms and subject headings based on the LCSH system.	Supporting: 15; 5:165–240.

Lester (1989) tested this assertion using twenty-two different retrieval processes on user queries to effect matches of subject headings. Of the many processes, she demonstrated that right truncation, string searching with adjacency, and keyword searching with an implicit Boolean "and" operator significantly improved match success.

Based on the results of an empirical study of user queries, Drabenstott and Vizine-Goetz (1994) concluded that the subject terms users entered into online systems possessed certain characteristics that revealed the subject searching approaches most likely to produce useful titles on the topics users sought. They suggested that computers systems could be programmed to identify many of these characteristics without the aid of human intermediaries, and, thus, respond to user queries with subject searching approaches likely to retrieve useful titles.

Generalization	Studies
19. Users enter queries containing language not used in subject headings in the LCSH system.	Supporting: 3:44; 2:41; 5:214–5.

Between two and five percent of user queries contained language not used in the LCSH system.

Generalization	Studies
20. The subject queries users enter into online systems can be used as *see* references to authorized subject headings.	Supporting: 22:68.

Designers of the experimental Okapi online catalog in the United Kingdom programmed Okapi to find certain words and phrases in user queries and search for them and their equivalent terms.

2.1.2 What Have We Learned about the Names Users Enter?

In *Using subject headings for online retrieval* (1994), Karen Drabenstott and Diane Vizine-Goetz summarized what had been learned to date about the names users enter into online catalogs in the form of five generalizations. This section gives brief discussions of generalizations. It cites studies that support or fail to support generalizations in the form of study numbers and page numbers; the former correspond to the studies listed in the "References" section that concludes this chapter. (The reference lists of studies that support or fail to support generalizations include a few more studies that have been published since the completion of *Using subject headings for online retrieval*.)

Generalizations were drawn from studies of name access points entered by users conducting searches for known-items. Findings about the personal, corporate, and geographic names users enter in known-item searches can be generalized to subject searches for names because many systems required users to enter name access points for known-item searches in the same form as name access points for subject searches in order to retrieve bibliographic records. Generalizations #2 about punctuation and #4 about misspelled words from the subjects section (section 2.1.1) were also true for name access points. Generalization #10 from the subjects section was true with a minor change, that is, users entered queries using the online catalog's personal name as author search capability that were probably better suited to other catalog capabilities.

Generalization	Studies
1. Users enter personal names in direct form when systems require the entry of inverted forms and vice versa.	Supporting: 4:32; 20:8; 8:51; 19:270; 12:400; 24:52; 11:27; 1:65, 67; 5:224.

Several studies drew this conclusion about the personal-name access points users entered into online catalogs. In Dickson's study (1984, 32), the percentage of personal names with given names entered first was as high as 37%.

Generalization	Studies
2. Users enter two personal names at the same time.	Supporting: 4:32; 20:8; 24:52; 11:27; 1:65; 5:225.

Users occasionally searched for material by or about more than one person. Examples of such queries were “berger, peter and luckman, thomas,” “leopold and loeb,” and “nietzche and kierkegard.”

Generalization	Studies
3. Users enter personal names with other words.	Supporting: 20:8; 15:186; 5:227–9.

Drabenstott and Vizine-Goetz (1994, 159, 227) reported that less than 1% of user queries contained elements for personal names and other words and gave examples: “paintings of pollack” [sic], “descartes future prediction,” and “clarence darrow’s relegious [sic] views.”

Generalization	Studies
4. The middle names or initials users include in their queries for personal names are sometimes counterproductive in helping them find the heading used in the catalog.	Supporting: 4:32; 23:80; 19:270; 12:400.

Of several reasons why the inclusion of middle names or initials was counterproductive, most centered on characteristics of the system into which the query was entered. For example, if the indexed heading had no initial and the user-entered term had an initial, the system placed the user in an alphabetical index of personal names following the name desired by the user and offered no capability for browsing backward in the alphabet.

Generalization	Studies
5. Retrieval processes in online catalogs can increase the match success rate between user-entered names and the access points indexed in bibliographic databases much more than the LCNAF.	Supporting: 20:15; 15:258; 5:230–3.

Taylor (1984, 15) and Lester (1989, 258) concluded that name authority records would help users find the authorized headings for names for only 6% and 1% of the user-entered terms in their respective studies. According to Taylor, a computer program to invert the elements in user-entered names that produced no retrievals would have helped users find an appropriate name heading for 22% of their access points. Both researchers recommended a right truncation program and Lester recommended string and keyword searching approaches for names. Drabenstott and Vizine-Goetz (1994, 230–3) took a different approach; they advocated that systems prompt users for elements of personal-name queries and submit queries to subject

searching approaches depending upon the ability of certain approaches to produce retrievals for query elements.

2.2Subject Searching Improvements

The objective of researchers who conducted studies of the subject terms users entered into online catalogs was to recommend improvements to the subject searching capabilities of online catalogs. Their calls for subject searching improvements were joined by other experts and spokespersons who were thinking about, commenting on, and analyzing research findings. In preparation for her dissertation research, Lester (1989, 58–98) reviewed recommendations for subject searching improvements made by both online catalog researchers and experts such as the following:

- Automatic detection and correction of spelling errors.
- Automatic detection and correction of search format errors.
- Automatic replacement of user queries with *see* references from LCSH.
- System-supplied truncation.
- Boolean-based, keyword searching of subject-rich fields of bibliographic records.

At the conclusion of her empirical study of the subject terms users enter into online catalogs, Lester (1989, 267) recommended “the retrieval processes of right truncation, string searching, and keyword searching” because “each makes significant improvements in match success with the Library of Congress subject headings” but she did not specify which retrieval process was better suited to which user queries.

Such lists of recommendations are limited for several reasons. They do not tell us which improvement will improve the prospects for success for the greatest number of queries entered into online catalogs. Lists do not tell us whether there are certain characteristics of user queries that make them better suited for particular subject searching improvements. They also do not tell us which subject searching capabilities in our existing online catalogs are currently working satisfactorily with user queries and the characteristics of these queries.

2.3 Subject Searching Improvements to Provide Useful Information

While previous studies increase our knowledge of user queries and provide several recommendations for improving subject searching in online catalogs, they leave a key question unanswered, viz. how can online systems respond to user queries with the subject searching approach most likely to succeed in providing relevant information?

The answer to this question was the objective of an empirical study of the subject terms users enter into online catalogs. The findings of this study were published in the research monograph *Using subject headings for online retrieval* (Drabenstott and Vizine-Goetz, 1994). Since the empirical study's findings were the impetus for the research described in this report, we felt it was important to highlist findings of the empirical study in this report. Findings are summarized in subsections of section 2.3.

2.3.1 Computer and Manual Analyses of Subject Queries

In the empirical study, the researchers extracted 54,429 subject queries entered by end users from the transaction logs of three online catalogs: (1) LS/2000 at the University of Kentucky, (2) ORION at UCLA, and (3) SULIRS at Syracuse University. They submitted extracted terms to a computer analysis that involved three successive comparisons to determine how closely user queries for subjects matched the established headings and *see* references in the machine-readable Library of Congress Subject Headings (LCSH-mr).

Almost 25% of subject queries were exact matches of established headings and *see* references in LCSH-mr. (Exact matches disregarded capitalization and punctuation.) Only one percent of subject queries matched normalized forms of established headings or *see* references in LCSH-mr. (Normalization disregarded capitalization, punctuation, stopwords, and word order in queries, and LCSH headings and *see* references.) About 14% of user queries were keyword matches of words in established heading and *see* reference fields of LCSH-mr records. (Keyword matches required every word in the subject query to match keywords in established heading and *see* reference fields of LCSH-mr records). When the three comparisons were performed and matches discarded, about 40% of user queries remained.

The computer analysis gave the researchers no opportunity to determine answers to qualitative questions such as the characteristics of remaining queries and the relevance of matching headings and *see* references. Thus, they conducted a manual analysis of subject terms to answer the following qualitative questions:

1. How closely does the query match the catalog's controlled vocabulary?
2. If the matching term is satisfactory for expressing the user's topic of interest, what system capability would quickly and efficiently deliver the user to the satisfactory term?
3. If the matching term is not satisfactory, what controlled vocabulary term is satisfactory for expressing the user's topic of interest and what system capability would quickly and efficiently deliver the user to the satisfactory term?

The researchers divided queries into groups for topical subjects, geographical names, corporate names, personal names, and various combinations of topical subjects and different types of names. They used a category scheme devised by Carlyle (1989) to determine whether the initial user queries in 1,503 subject searches were exact, partial, keyword, or nonmatches of the established headings and *see* references in LCSH-mr.

2.3.2 Key Findings of the Empirical Study

Five key findings of the empirical study of user queries are enumerated below along with system capabilities that were supported by the particular finding.

1. Topical subject queries and geographical name queries register high percentages of exact matches.

The matching term often was posted in hundreds of bibliographic records. For example, user queries such as "art," "computers," "photographs," "united states," and "great britain" were exact matches of LCSH-mr headings that retrieved hundreds of bibliographic records in many library databases. The number of retrieved bibliographic records would increase into the thousands if subdivided forms of these headings were retrieved.

To perform retrieval quickly and efficiently, online catalogs need a variation of the alphabetical approach to subject searching. Such a variation would anticipate the user's selection of the exact match from an alphabetical list, and thus begin with a report of the results of such exact matches, summarize the retrieval of subdivided forms, and give users the option to browse related terms from LCSH-mr if available.

2. Partial matches are: (1) most combinations of queries for topical subjects and names, and (2) between one-quarter and one-third of queries for topical subjects and geographical names, respectively.

Partial matches were satisfactory representations of the topics of interest users had in mind only when queries matched the initial words in a longer assigned heading or LCSH-mr *see* reference. The alphabetical approach would be helpful for placing users in the neighborhood of potentially relevant controlled vocabulary terms. Different subject searching approaches would be required for other partial matches.

3. Most user queries would be satisfied by the alphabetical or keyword approaches implemented in existing online catalogs.

Existing online catalogs feature the following subject searching approaches: (1) alphabetical, (2) keyword-in-main-heading, (3) keyword-in-subdivided heading, and (4) keyword-in-record. Except for a few queries, most queries would retrieve useful material using one or several of these approaches. Although many online catalogs feature more than one of these approaches, they do not provide the user with guidance as to which approach will provide useful results in response to a particular query.

4. The subject terms users enter into online systems possess certain characteristics that reveal the subject searching approaches most likely to succeed at providing useful information on the topics users seek.

Examples of these characteristics were the number of words in user queries, the extent to which user queries matched controlled vocabulary terms, and their ability to produce retrievals in response to certain subject searching approaches. Online systems could help users choose among existing subject searching approaches by determining how closely their subject queries match assigned subject headings in library databases and produce retrievals.

5. Users enter the various elements of personal name queries in an unpredictable sequence.

The order in which users entered the elements of personal name queries was unpredictable even when users entered personal name queries in systems that required a particular form of name, e.g., direct form or inverted form. Systems need a separate subject search option for personal names. When users select such an option, systems should prompt users to enter the four elements of personal name queries, i.e., last name, first name, middle name or initial, and topic, or as many of the elements as they know. The particular elements users provide and the system's ability to find these elements in personal name headings determines the subject searching approaches used by the system.

2.4Subject Searching Functionality and Search Trees

2.4.1 Subject Searching in Operational Online Catalogs

Operational online catalogs feature keyword and alphabetical approaches to the subject queries users enter. The former approach gives a report of the number of assigned subject headings or cataloging records bearing words in the user query. The latter approach produces a list of assigned subject headings in the alphabetical neighborhood of the user query.

Consider the user query "aids." When "aids" is entered through keyword approaches, systems retrieve thousands of cataloging records but most retrieved records are not relevant because retrievals are made for the word "aids" in the subdivision "Audio-visual aids" that is connected to hundreds of subject headings on many different topics. Results of the alphabetical approach would be much more useful because the list of subject headings in alphabetical neighborhood of "aids" would probably include the relevant heading "AIDS (Disease)" and several subdivided forms of this heading, e.g., "AIDS (Disease) — Alternative treatment" and "AIDS (Disease) — Biography." Although both approaches retrieve relevant cataloging records, most records retrieved through the alphabetical approach are going to be specifically about the AIDS disease because they are retrieved through the relevant subject heading for this topic.

How would online catalog users know whether to use keyword or alphabetical approaches for this query? Do users know the differences between approaches? Although the majority of online catalogs in American libraries offer more than one approach to subject searching, they give users little, if any, guidance as to which approach is better suited to their queries. Some catalogs feature as many as five subject searching approaches! Users could enter their queries using several approaches and evaluate the results for themselves. But how many users would be patient enough to review the results of more than one subject searching approach to determine the most appropriate system response?

2.4.2 Search Trees: An Innovative Approach to System Design

Search trees hold much promise for assuming the burden of determining which subject searching approach is likely to produce useful information for user queries. The designers of the Okapi experimental online catalog first defined search trees as "a set of paths with branches or choices, which enables the system to carry out the most sensible search function at each stage of the search" (Mitev, Venner, and Walker 1985, 94).

The search trees they implemented in OKAPI "evolved through a process of discussion and trial and error" and placed more emphasis on searching the titles than the subject headings in OKAPI's cataloging records because only half of these records contained subject headings (Mitev, Venner, and Walker 1985, 94).

Some online catalogs have subject searching routines that resemble search trees. For example, the online catalog of the University of Illinois at Urbana-Champaign responds to user queries for subjects with keyword searches of assigned subject headings. When users terminate searches, the system prompts them to continue and gives the results of a title-keyword search (Hildreth 1989, 86–7). The Illinois online catalog always performs keyword searches of subject heading fields before title-keyword searches because the former consumes fewer system resources than the latter.

The search trees tested in this research effort were the result of the empirical study of the subject terms users entered into online catalogs (Drabenstott and Vizine-Goetz 1994). These search trees emphasized subject headings because the vast majority of cataloging records created by American libraries are assigned subject headings based on the Library of Congress Subject Headings (LCSH) (O'Neill and Aluri 1979, 5).

The search trees exemplified the searching strategies used by expert search intermediaries. Intermediaries use controlled vocabulary because it yields relevant output. When controlled vocabulary is not available to express user queries, they conduct free text searches of titles and abstracts to retrieve a few relevant records, review results to find relevant controlled vocabulary, and then incorporate such vocabulary into the ongoing search. The search trees performed in a similar manner. They invoked searching approaches that looked for matches of user queries in subject heading fields of cataloging records before enlisting keyword search approaches that looked for matches in title fields or in a combination of title and subject heading fields. Chapter 4 gives a full description of the search trees tested in this study.

2.5Chapter Summary

This chapter presented generalizations about user queries for subjects. Generalizations were derived from studies of user queries in which researchers sampled user queries from transaction logs or selected user queries from a transaction log recorded over a certain time period.

Although previous studies increase our knowledge of user queries and provide several recommendations for improving subject searching in online catalogs, they leave a key question unanswered, viz. how can online systems respond to user queries with the subject searching approach most likely to succeed in providing relevant information?

The answer to this question was the objective of an empirical study of the subject terms users enter into online catalogs (Drabenstott and Vizine-Goetz 1994).

The findings of the empirical study were the foundation for the new design of subject access to online catalogs that included search trees. Search trees are a set of paths with branches or choices that enable systems carry out the most sensible search function at each stage of the search. The search trees tested in this study exemplified the searching strategies used by expert search intermediaries who favor searches of controlled vocabulary fields over other content-rich fields of cataloging records. Search trees hold much promise for assuming the burden of determining which subject searching approach is likely to produce useful information for user queries. This report describes a test of the new subject access design in the ASTUTE experimental online catalog.

References

1. Ballard, Terry, and Jim Smith. 1992. "The human interface: an ongoing study of OPAC usage at Adelphi University." In *Advances in online public access catalogs*, edited by Marsha Ra, 58–73. Westport, CT: Meckler.

2. Barrett, Beverly, and Margaret Maticka. 1990. "An analysis of user failure in subject searching an online catalogue." In *Garbage in, garbage out: The need for quality in the age of automation; Australian Library and Information Association 8th national cataloguing conference*, 38–49. Adelaide: Auslib Press.

3. Carlyle, Allyson. 1989. "Matching LCSH and user vocabulary in the library catalog." *Cataloging & Classification Quarterly* 10, 1/2: 37–63.

4. Dickson, Jean. 1984. "An analysis of user errors in searching an online catalog." *Cataloging & Classification Quarterly* 4, 3 (Spring): 19–38.

5. Drabenstott, Karen M., and Diane Vizine-Goetz. 1994. *Using subject headings for online retrieval: Theory, practice, and potential*. San Diego: Academic Press.

6. Freiburger, Gary, and Marjorie Simon. 1983. "Patron use of an online catalog." In *Crossroads: Proceedings of the first national conference of the Library and Information Technology Association*, edited by Michael Gorman, 106–11. Chicago: American Library Association.

7. Hancock, Micheline, et al. 1990. *Evaluation of online catalogues: an assessment of methods*. London: British Library. British Library Research Paper no. 78.

8. Henty, Margaret. 1986. "The users at the online catalogue: a record of unsuccessful keyword searches." *LASIE* 17, 2 (September/October): 4–52.

9. Hildreth, Charles R. 1989. *Intelligent interfaces and retrieval methods for subject searching in bibliographic retrieval systems*. Washington, DC: Library of Congress. Advances in Library Information Technology 2.

10. Hildreth, Charles R. 1985. "Monitoring and analyzing online catalog user activity." *LS/2000 Communiqué*, pp. 3–6.

11. Holmes, David, and Derrick Bulger. 1988. "A day in the life of a public terminal — a transaction analysis of an online catalogue terminal in a bilingual environment." *Canadian Journal for Information Science*, 13, 3/4: 21–33.

12. Hunter, Rhonda N. 1991. “Successes and failures of patrons searching the online catalog at a large academic library: a transaction log analysis.” *RQ* 30, 3 (Spring): 395–402.

13. Jones, Richard. 1986. “Improving Okapi: transaction log analysis of failed searches in an online catalogue.” *Vine* no. 62: 3–13.

14. Kurth, Martin. 1994. “The limits and limitations of transaction log analysis.” *Library Hi Tech* 11, 2: 98–104.

15. Lester, Marilyn Ann. 1989. *Coincidence of user vocabulary and Library of Congress Subject Headings: experiments to improve subject access in academic library online catalogs*. Ph.D. dissertation, University of Illinois at Urbana-Champaign.

16. Markey, Karen. 1984. *Subject searching in library catalogs: before and after the introduction of online catalogs*. Dublin, Ohio: OCLC.

17. Mitev, Nathalie, Gillian Venner, and Stephen Walker. 1985. *Designing an online public access catalog*. London: British Library. Library and Information Research Report 39.

18. O’Neill, Edward T., and Rao Aluri. 1979. *Subject heading patterns in OCLC monographic records*. Columbus, Ohio: OCLC, Inc. OCLC Research Report Series OCLC/OPR/RR–79/1.

19. Peters, Thomas A. 1989. “When smart people fail: An analysis of the transaction log of an online public access catalog.” *Journal of Academic Librarianship* 15, 5 (November): 267–73.

20. Taylor, Arlene G. 1984. “Authority files in online catalogs: An investigation of their value.” *Cataloging & Classification Quarterly* 4, 3 (Spring): 1–19.

21. Van Pulis, Noelle, and Lorene E. Ludy. 1988. “Subject searching in an online catalog with authority control.” *College & Research Libraries* 49 (6): 523–33.

22. Walker, Stephen, and Richard M. Jones. 1987. *Improving subject retrieval in online catalogues; 1. Stemming, automatic spelling correction and cross-reference tables*. London: British Library. British Library Research Paper no. 24.

23. Walter, Dennis R. 1987. “The user at the online catalogue: A record of unsuccessful keyword searches — another case study.” *LASIE* 18, 3 (November/December): 74–81.

24. Zink, Steven D. 1991. “Monitoring user search success through transaction log analysis: the WolfPAC example.” *Reference Services Review* 19, 1 (Spring): 49–56.

3 Machine-readable Data for System Development

3.1 Introduction

Chapters 3–5 highlight the development of the ASTUTE experimental online catalog. This chapter begins with a description of the computer equipment used for ASTUTE development. It features a description of the machine-readable bibliographic and authority data that serve as the foundation for ASTUTE's searchable databases and dictionaries.

3.2Computer Equipment

ASTUTE (A Search Tree Underlying the Experiment) was programmed on a stand-alone Gateway 2000 486, 33 MHz, IBM-compatible microcomputer, with 8 megabytes of RAM and a VGA color monitor. The operating system was MS-DOS version 5.0. A dot-matrix printer and a mouse were attached to the microcomputer for use by ASTUTE project staff during development work and end users during online retrieval tests.

The ASTUTE project team used this computer equipment and a connection to the university' campus network for various tasks, e.g., transmitting data on magnetic tape to the microcomputer's hard disk, transmitting transaction log files of end user search activity between participating institutions. When ASTUTE was installed in the libraries of University of Michigan-Dearborn (UM-D) and Earlham College, the system did not require a network connection. It resided entirely on the Gateway microcomputer. Participating library staff monitored the system, performed daily system backups, and periodically used the microcomputer equipment and network connections in their institutions to transmit transaction log files of end user search activity to the ASTUTE project team in Ann Arbor.

3.3 Machine-readable Bibliographic Data

The databases of the ASTUTE experimental online catalog were created from two data sources: (1) machine-readable cataloging (MARC) records for bibliographic data from the two participating libraries in selected subject areas of the Library of Congress Classification (LCC), and (2) MARC records for subject authority data from the compact disk-based product CD/MARC Subjects distributed by the Library of Congress. The number and subject areas of MARC bibliographic records were:

1. Mardigian Library of the University of Michigan-Dearborn: 14,686 bibliographic records in Computer Science (QA76's) and Technology (T–TX).
2. Lilly Library of Earlham College: 11,976 bibliographic records in American History (E1–F1199).

The ASTUTE project team did not combine bibliographic records into a single database. Rather, the team used the two libraries' bibliographic records to create separate, searchable databases on computer science and technology for UM-D and on American history for Earlham College, respectively. During online retrieval tests at Mardigian Library and Lilly Library, users searched a database of materials available at the library at which they were searching ASTUTE.

3.3.1 Obtaining Bibliographic Data from Participating Libraries

UM-D provided bibliographic records in USMARC format on two magnetic tapes in ASCII format. To transfer records from tape to the Gateway microcomputer's hard disk, the tape was mounted at the University of Michigan (UM) computer center in Ann Arbor, copied to a file on the mainframe, and the file was downloaded to the Gateway microcomputer's hard disk. All UM-D bibliographic records were accepted for processing into ASTUTE.

An ASTUTE project team member traveled to Earlham College to download USMARC format bibliographic records from Lilly Library's Marcive Public Access Catalog (PAC) to floppy disk. The team member used Marcive PAC's call number browsing capability to produce brief title displays in call number order, highlight titles in blocks of two hundred, and, for highlighted titles, download USMARC-format bibliographic records to floppy disk in ASCII format. Downloaded records on floppy disk were loaded directly into the Gateway microcomputer in the project team office in Ann Arbor. Downloaded records included duplicate records because of the task-intensive manual downloading process; consequently, the project team included a

routine in the conversion program described in section 3.3.2 to avoid writing duplicate records. Except for duplicates, all Earlham records were accepted for processing into ASTUTE. Since bibliographic records from UM-D and Earlham were USMARC-format records, the ASTUTE project team used similar computer programs to process bibliographic data into ASTUTE for database building.

3.3.2 Converting Bibliographic Data

After ASCII-format bibliographic records were loaded into the Gateway microcomputer's hard disk, the ASTUTE Project Team wrote a bibliographic conversion program to convert ASCII-format bibliographic records contributed by UM-D and Earlham into databases (or files) for local processing format. The conversion programs were written using low-level file access functions in FoxPro. The programs created seven FoxPro database tables and two temporary databases for listing subject headings (USMARC tags 650 and 651). Additional database tables were created from the original seven databases with several programs, depending on the search criteria of the individual search trees.

Programs written for converting UM-D and Earlham records were very similar. The only differences were connected with parameters of Library of Congress call number fields and the ASCII characters beginning each bibliographic record in the file.

USMARC-format records for bibliographic data consist of a 24-character leader (or header), a directory record bearing a series of twelve characters for each tag within the record that provide the map for finding data fields, and the variable-length data fields in the record. Conversion programs first processed the directory records for each bibliographic record. They skipped the leader in each record (i.e. the first twenty-four characters) and positioned the file pointer at the beginning of directory records. Each directory record was read in individually twelve characters at a time into a memory variable. Programs checked the first character of the memory variable to see if the character was ASCII character 30 which indicated the end of the directory. If the end of the directory was not found, programs divided directory records into the USMARC tag number, the length of the record, and the starting position of the record and these were stored in a directory array or table in memory. The program continued reading and storing directory records into an array until ASCII character 30 was found indicating the end of the directory. When ASCII character 30 was found, the file pointer was set back eleven characters to the beginning of the variable length data fields. The program then sorted the table of directory records on the order of the starting position of the variable-length fields in the file.

Conversion programs processed variable-length fields in bibliographic records after the directory records were stored in the directory array. They checked the USMARC tag in the array to see if the tag was to be converted into the FoxPro database. If the field was needed, the file pointer was placed at the starting position in the file of the variable-length field, and the total number of characters for the record were read into a memory variable. Conversion programs searched for ASCII character 31 which indicated subfield divisions within a field. Conversion programs assigned each bibliographic record a unique number to index on a unique field for data lookups. Programs distributed bibliographic record data in desired fields and subfields into five FoxPro bibliographic databases and the two temporary databases for subject headings bearing USMARC tags 650 and 651, respectively. Programs continued processing each element in the directory array until all of the USMARC fields in the directory array were written to the appropriate FoxPro databases.

Conversion programs then continued to read in data one character at a time to find the beginning of the next record. A check was done to see if the end of the file had been found. If this check found the end of the file, processing was finished and if not, the character was checked to see if it was the beginning of the next bibliographic record. ASCII character 29 was the beginning of each UM-D record and ASCII character 10 was the beginning of each Earlham record. When the beginning of the new record was found, the program again skipped the twenty-four character header and continued processing directory records into a new directory array.

The project team designed the bibliographic databases and tested the conversion programs thoroughly to make sure that accurate information was imported into the FoxPro databases on the Gateway microcomputer.

3.3.3 Changing Bibliographic Data

3.3.3.1 Diacritics

ASTUTE was developed using the FoxPro database management system version 2.0 (FoxPro Software, Inc. 1991). FoxPro supports and displays the standard 255 ASCII character set. The standard ASCII characters for many of the diacritics found in the bibliographic records were not displayed as they would be on a system using a different character set. When FoxPro displayed bibliographic data bearing diacritics, it indicated the presence of diacritics by displaying a standard ASCII character in the ASCII 255-character set (i.e., the accent mark sometimes printed above the last “e” in the word resume was displayed as a separate ASCII character). The ASTUTE project team did not delete diacritics from UM-D bibliographic records because diacritics

were so uncommon in the UM-D database of computer science and technology records that their presence would not adversely affect retrieval.

Diacritics were much more common in Earlham College bibliographic records on American history. Since the presence of diacritics could adversely affect retrieval, the ASTUTE Project Team wrote, tested, and executed a program that deleted diacritics from bibliographic records.

3.3.3.2 Subject Headings

Mardigian Library (UM-D) staff advised the ASTUTE project team that the subject headings in their library's bibliographic records would contain errors. The project team wrote, tested, and executed a program to produce two alphabetical lists of unique subject headings for topics and geographic names (USMARC tags 650 and 651, respectively). In the process of downloading authority records from CD/MARC Subjects (see section 3.3.4), project team reviewed listed subject headings and annotated the list with corrections. The project team manually corrected erroneous subject headings in UM-D bibliographic records. Frequent errors in main headings and subdivisions were: (1) assignment of *see* references, (2) incorrect tags, (3) incorrect subfield codes, (4) misspellings, (5) abbreviations, (6) canceled subject headings, and (7) canceled subdivisions.

Although Lilly Library (Earlham) staff did not expect many erroneous subject headings, ASTUTE project staff also manually corrected erroneous subject headings in Earlham bibliographic records. Since these records contained many names used as subjects, the project team produced a third alphabetical list for unique names as subjects (USMARC tag 600) and encountered errors they did not encounter with topical subjects, e.g., incorrect dates, open date ranges, and missing qualified names.

3.3.4 Enhancing Bibliographic Data

3.3.4.1 Form Subdivisions

The objective of ASTUTE's exact approach to subject searching (see section 5.3) was to manage lengthy browsing displays under a given search term using broad categories. Broad categories corresponded to subfield codes in subject heading fields, thus, they were limited to three subfield codes: (1) $x for topical subdivisions, (2) $y for period subdivisions, and (3) $z for geographic subdivisions.

The project director and others have called for an editorial review of subdivisions in the *Subject cataloging manual: subject headings* (SCM:SH) for the purpose of creating broad categories (Massicotte 1988; Holly and Killheffer 1982; Drabenstott and Vizine-

Goetz 1990, 13, 1994, 254–60). The editorial review would be a huge undertaking because of the thousands of subdivisions in SCM:SH. In the absence of such a review, the ASTUTE project team created one new broad category for form subdivisions. The team reviewed an alphabetical list of unique topical subdivisions in UM-D and Earlham bibliographic records and reassigned selected subdivisions to a new "local" subfield code ($1). Examples of form subdivisions are "Amateurs' manuals," "Bibliography," "Films," "Maps," and "Study guides." Appendix A gives two separate lists of the form subdivisions in UM-D and Earlham bibliographic records.

3.3.4.2 Control Field Codes

Several coded positions in the 008-control field of USMARC bibliographic records convey information about the physical and bibliographic characteristics of the item. Unfortunately, replacing the codes with control-field codes has been ignored or overlooked in cataloging and in system design (Byrne (1987a, 4).

Realizing the value of coded data in the 008-control field, the ASTUTE project team identified the following four control field elements that would be useful in subject searches and bibliographic record displays: (1) illustration codes (positions 18–21), (2) nature of contents codes (positions 24–27), (3) biography code (position 34), and (4) festschrift code (position 30). The team programmed ASTUTE to convert codes in these control-field positions to the English-language equivalent terms for use in subject searching and displaying bibliographic records. ASTUTE includes English-language equivalent terms from these four elements in keyword-in-record searches for subjects generally and for personal names as subjects. For example, a keyword-in-record search in ASTUTE for the user query "maps of us 40" produced retrievals from several subject-rich fields including the 008-control field. One bibliographic record that was retrieved in the search was *US 40: a roadscape of the American experience*. Figure 3.1 shows this record; it bears the terms "us" and "40" in the title field and "maps" in the "other contents" field.

Appendix B lists English-language equivalent terms and display field labels for the four control field elements ASTUTE used in keyword-in-record searches for subjects generally and for personal names.

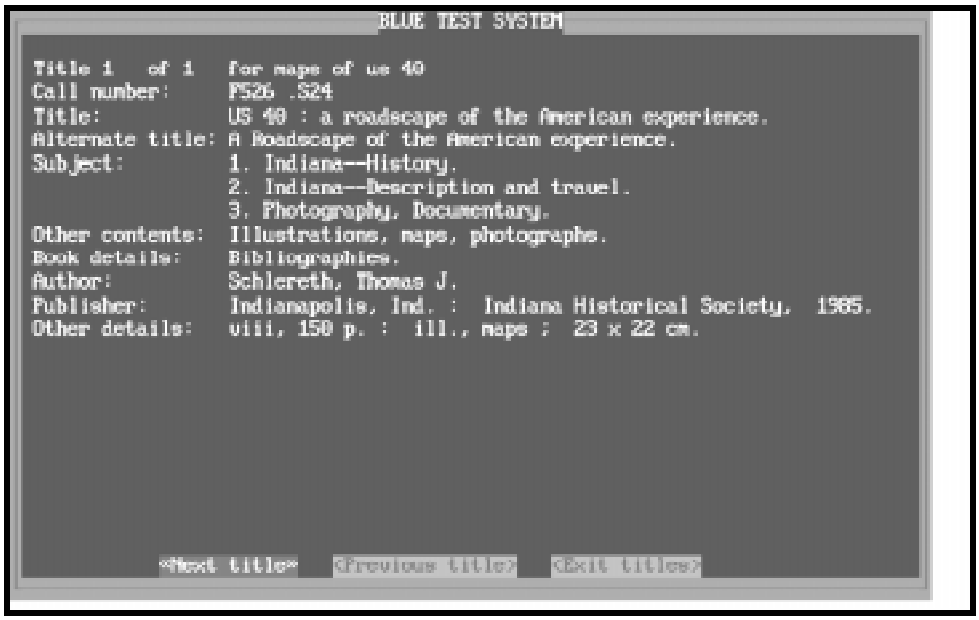

Figure 3.1. Title bearing terms for control-field codes

3.4Subject Authority Data

3.4.1 Obtaining Subject Authority Data

In 1986, the Library of Congress (LC) began distributing subject authority data to subscribing libraries. Libraries responded by requiring subject authority data functionality in system specifications and setting up operations to maintain the currency of subject authority data in their automated systems (Garrison 1991). Vendors of automated library systems responded by enhancing their online catalogs with subject authority data and functionality to handle such data (Drabenstott 1991, 5–6).

Subject authority records have several features to help subject searchers: (1) scope notes to define the use of the subject heading, (2) broader, narrower, and related terms to help users refine their topics of interest, and (3) cross references to guide users from unused words and phrases to authorized subject headings. The ASTUTE project team added subject authority data and functionality into ASTUTE because library staff and catalog users take these features for granted in today's online catalogs.

LC distributes machine-readable subject authority data on magnetic tapes (machine-readable Library of Congress Subject Headings [LCSH-mr]) and on compact disk (CD/MARC Subjects). The ASTUTE project team inquired about the availability of

LCSH-mr or CD/MARC Subjects. LC responded by giving the project team a complimentary copy of CD/MARC Subjects.

The ASTUTE project team programmed FoxPro to generate alphabetical lists of unique subject headings for topical subjects and geographic names (MARC tags 650 and 651) in UM-D and Earlham bibliographic records. Team members used these lists to search CD/MARC Subjects. They downloaded subject authority records for subject headings that matched the main subject heading or the main subject heading and one or more subdivisions of listed headings. Table 3.1 cites the number of subject authority records selected and downloaded. The number of matching authority records was smaller than the number of unique subject headings because many subject headings were subdivided, and, thus, had no corresponding records in the subject authority file. Records downloaded from CD/MARC Subjects were USMARC format records for authority data. Records were loaded directly into the Gateway microcomputer in the project team office in Ann Arbor.

Table 3.1. Authority Records Selected

Library	Number of Unique Subject Headings	Number of Matching Authority Records
UM-D	7,792	4,219
Earlham	9,345	3,824

3.4.2 Converting Subject Authority Data

The ASTUTE project team wrote a conversion program to convert the USMARC authority records from ASCII format files to the FoxPro database format. The program contained a routine to avoid writing duplicate records. The conversion program converted the selected subject authority records to five FoxPro databases (or files). The conversion process was very similar to the one used to convert the bibliographic records (section 3.3.2). The ASTUTE project team handled diacritics in the same way as they handled diacritics in bibliographic records (section 3.3.3.1).

3.4.3 Enhancing Subject Authority Data with Narrower Terms

The ASTUTE project team wrote and implemented a program to verify whether broader and related terms in subject authority records were used in participating library bibliographic records. Unused terms were deleted from subject authority records. When broader terms were used in bibliographic records, they were added as narrower terms to the appropriate subject authority record. For example, the subject heading “Archaeology” was a broader term in the subject authority record for “Cliff-

dwellings." Since both terms were used in Earlham bibliographic records, the program added "Cliff dwellings" as a narrower term to the "Archaeology" authority record.

3.5Chapter Summary

Chapter 3 was the first of three chapters highlighting the development of the ASTUTE experimental online catalog. This chapter began with a description of the computer equipment used for ASTUTE development. It featured a description of the machine-readable bibliographic and authority data that served as the foundation for ASTUTE's searchable databases.

The databases of the ASTUTE experimental online catalog were created from two data sources: (1) machine-readable cataloging (MARC) records for bibliographic data from the two participating libraries in selected LCC subject areas, and (2) USMARC records for subject authority data from the compact disk-based product CD/MARC Subjects distributed by the Library of Congress. Details on how the ASTUTE project team obtained bibliographic data (section 3.3) and authority data (section 3.4) were provided.

The team made few changes to bibliographic data: (1) deleting diacritics in Earlham bibliographic records (section 3.3.3.1), and (2) correcting erroneous subject headings (section 3.3.3.2). The few enhancements the team made to bibliographic data were intended to improve certain subject searching approaches: (1) identifying and coding form subdivisions to increase the number of broad categories in exact search from three to four (section 3.3.4.1), and (2) converting control-field codes to English-language equivalent terms to increase the searchable dictionary in keyword-in-record searches (section 3.3.4.2). The team deleted diacritics in subject authority records (section 3.4.2) and enhanced these records with narrower terms (section 3.4.3).

References

Byrne, Deborah. 1987a. "The much-misunderstood MARC fixed field: Part 1 of 3." *Action for Libraries* (February): 4.

Byrne, Deborah. 1987b. "The much-misunderstood MARC fixed field: Part 3 of 3." *Action for Libraries* (April): 3–4.

Drabenstott, Karen Markey. 1991. "Introduction." In *Subject authorities in the online environment*, 1–8. Chicago: American Library Association.

Drabenstott, Karen M., and Diane Vizine-Goetz. 1994. *Using subject headings for online retrieval: Theory, practice, and potential*. San Diego: Academic Press.

Drabenstott, Karen Markey, and Diane Vizine-Goetz. 1990. "Search trees for subject searching in online catalogs." *Library Hi Tech* 8, 3: 7–20.

FoxPro Software, Inc. 1991. *FoxPro commands and functions*. Perrysburg, Ohio: FoxPro Software, Inc.

Garrison, William A. 1991. "Practical considerations in using the machine-readable LCSH." In *Subject authorities in the online environment*, 41–55. Chicago: American Library Association.

Holley, Robert P., and Robert E. Killheffer. 1982. "Is there an answer to the subject access crisis?" *Cataloging & Classification Quarterly* 1, 2/3: 125–33.

Massicotte, Mia. 1988. "Improved browsable displays for online subject access." *Information Technology and Libraries* 7, 4 (December): 373–80.

Karen M. Drabenstott and Marjorie S. Weller

4 Search Trees for Subject Searching

4.1 Introduction

The subject terms users enter into online systems possess certain characteristics that reveal the subject searching approaches most likely to succeed in terms of producing assigned subject headings and bibliographic records on the topics users seek. Computer systems can be programmed to identify many of these characteristics without the aid of a human intermediary. Examples of characteristics are the number of words in user queries, the extent to which user queries match controlled vocabulary terms, and their ability to produce retrievals in response to certain subject searching approaches. Programming computer systems to identify these characteristics would enable systems to respond to the subject queries users enter with the subject searching approaches most likely to succeed at providing relevant information.

The ASTUTE experimental online catalog was composed of two catalogs: (1) Blue Test System in which search trees govern the system's selection of subject searching approaches, and (2) Pinstripe Test System in which subject searching approaches were selected randomly. This chapter describes the search trees of the Blue System that control its selection of subject searching approaches (sections 4.2–4.5). It sets the stage for chapter 5 discussions of subject searching approaches in the Blue and Pinstripe Systems.

4.2 Search Tree Design

Search trees governed the Blue Test System's selection of subject searching approaches. In the Blue System, users did not explicitly choose a particular approach. Rather, the system responded with an approach based on the characteristics of user queries and their ability to produce retrievals. When the system selected an approach yielding zero or too few retrievals, users could continue searching using another approach. Search trees also governed the Blue System's selection of subsequent subject searching approaches.

Six search trees for subject searching are presented in this chapter. These search trees were implemented into the Blue Test System. Search trees incorporated into the Blue Test System emanated from the findings of an empirical study of the subject terms users enter into online catalogs (Drabenstott and Vizine-Goetz, 1994).

4.3 Initial Search Tree

The user's selection of the subject searching option in the Blue Test System set into motion the initial search tree. The initial search tree was a filter because it dispatched user queries to a particular search tree that favored the selection of certain subject searching approaches over others. Figure 4.1 is the initial search tree.

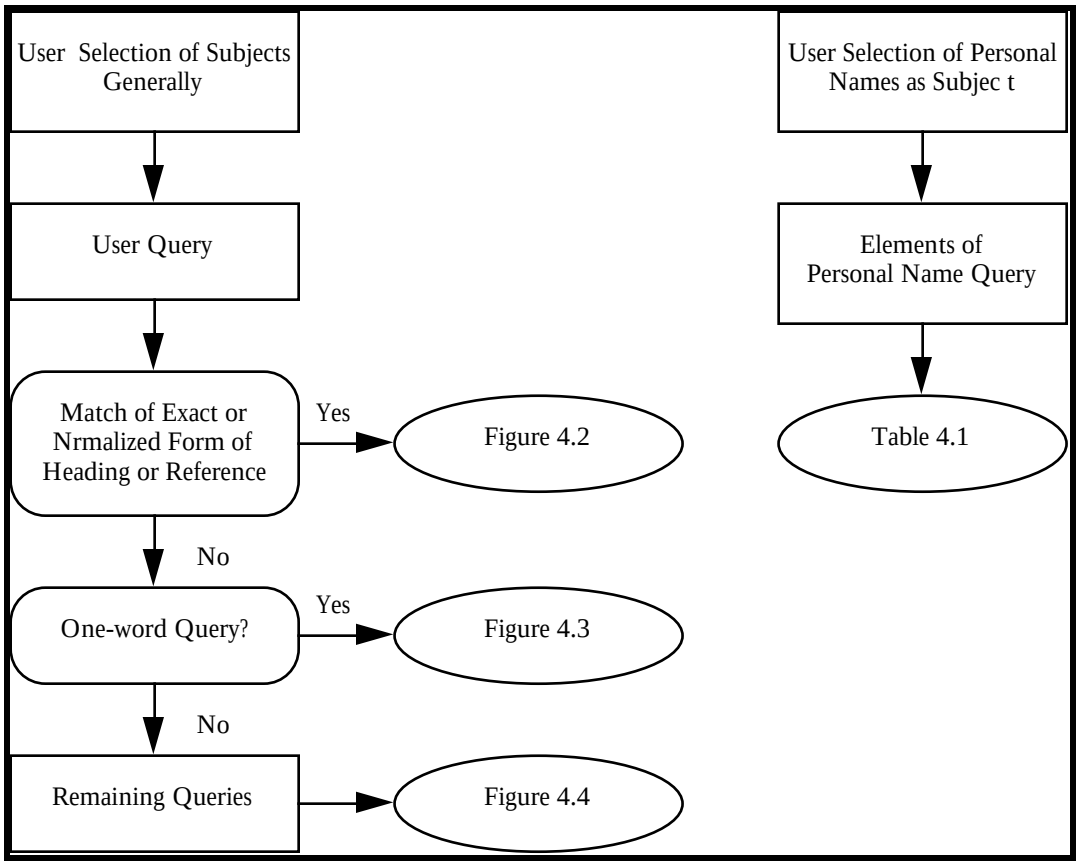

Fig. 4.1. Initial search tree

The initial search tree let users distinguish their queries for personal subjects from queries for topical subjects generally and, based on summary characteristics that systems determined about the latter types of user queries, dispatched them to a particular search tree that favored certain subject searching approaches over others. All user

queries for subjects generally were candidates for the exact approach. To effect an exact match, the Blue System manipulated user queries in the same way as controlled vocabulary terms were manipulated to establish exact and normalized forms, e.g., ignoring capitalization, removing punctuation and stopwords. In the event an exact match was found, the initial search tree dispatched the query to the search tree that governed the exact approach (figure 4.2).

Unable to find an exact match, the initial search tree dispatched user queries to one of two search trees based on the number of words in queries. One-word queries were given to a search tree that favored the alphabetical approach (figure 4.3). The extent to which remaining queries matched controlled vocabulary terms determined whether they were submitted to a search tree that favored the alphabetical or keyword approaches (figures 4.4 and 4.5, respectively).

The Blue System prompted users entering personal-name queries for the name and topic elements of personal-name queries. Personal-name queries were handled by a search tree that was separate from search trees for subjects generally (Table 4.1). The Blue System chose between alphabetical and keyword approaches depending on the types of elements users entered and the ability of these elements to produce retrievals.

4.4 Search Trees for Subject Queries Generally

The Blue System responded to user queries for subjects generally that matched exact or normalized forms of controlled vocabulary terms with the exact approach. Figures 4.2A and 4.2B depict the search tree that featured the exact approach. Details about the exact approach are given in chapter 5.

The exact approach search tree was split into two parts. Figure 4.2A depicts major events of the exact approach in which the system presented a conceptual map to subdivided forms of the matched heading, and, if available, gave users the option to browse related terms and other information about the matched heading.

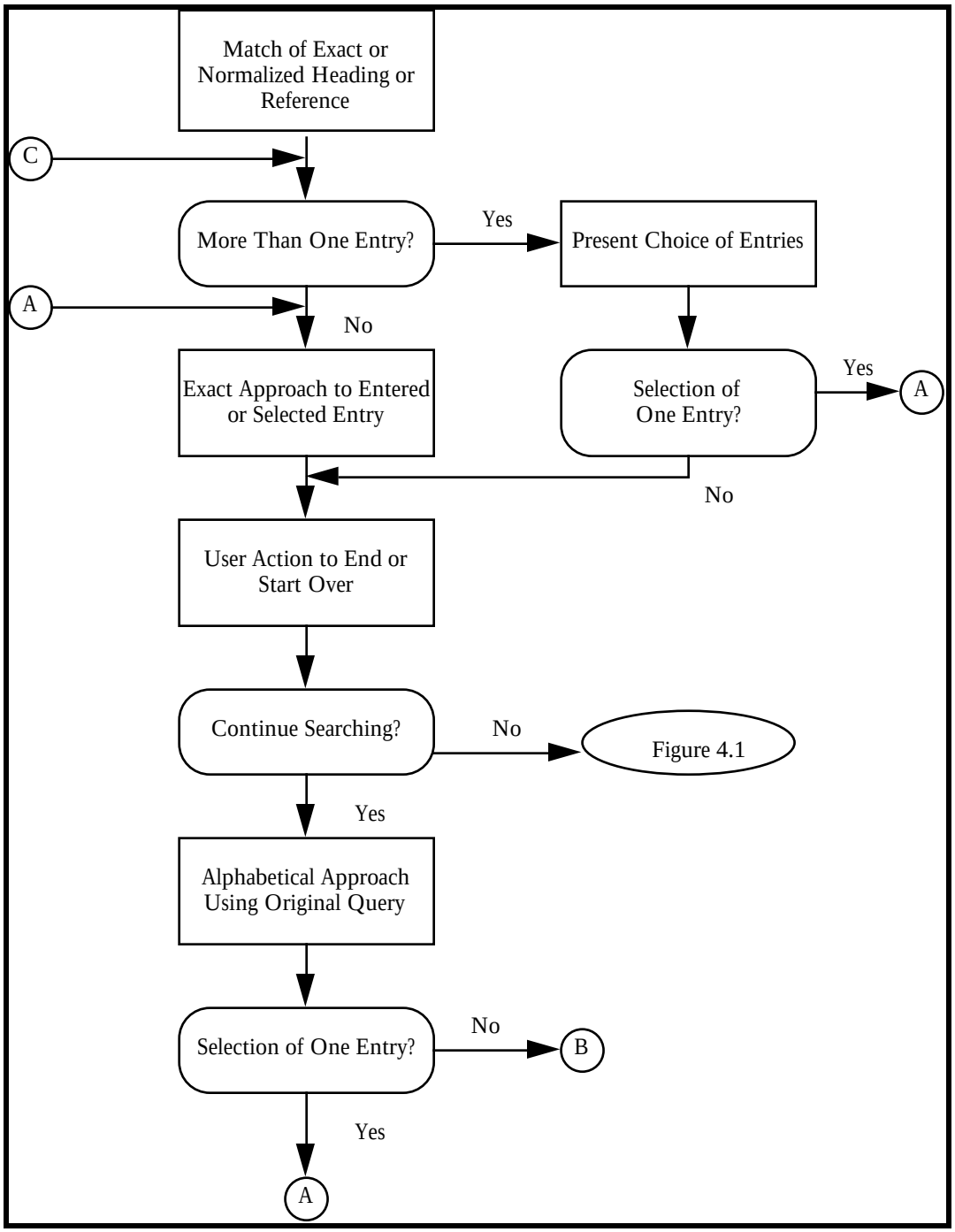

Figure 4.2A. Search tree for the exact approach

Figure 4.2B depicts major events following a user action to start over or end the search. The search tree submitted or the original user query to other approaches beginning with controlled vocabulary approaches. The Blue System continued searching using the

alphabetical approach and various keyword approaches beginning with the keyword-in-main-heading search.

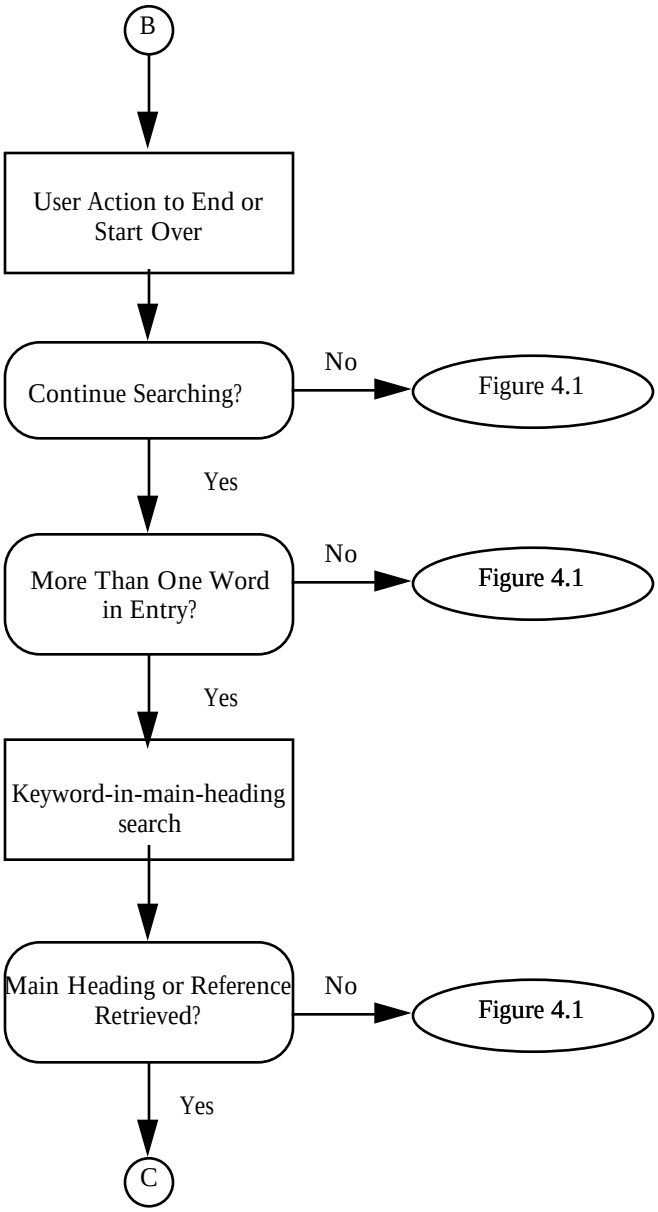

Figure 4.2B. Search tree for the exact approach (contd.)

Figure 4.2B shows only the keyword-in-main-heading search. This tree could be expanded to include keyword-in-subdivided-heading, title-keyword, keyword in subject heading fields, and keyword-in-record searches.

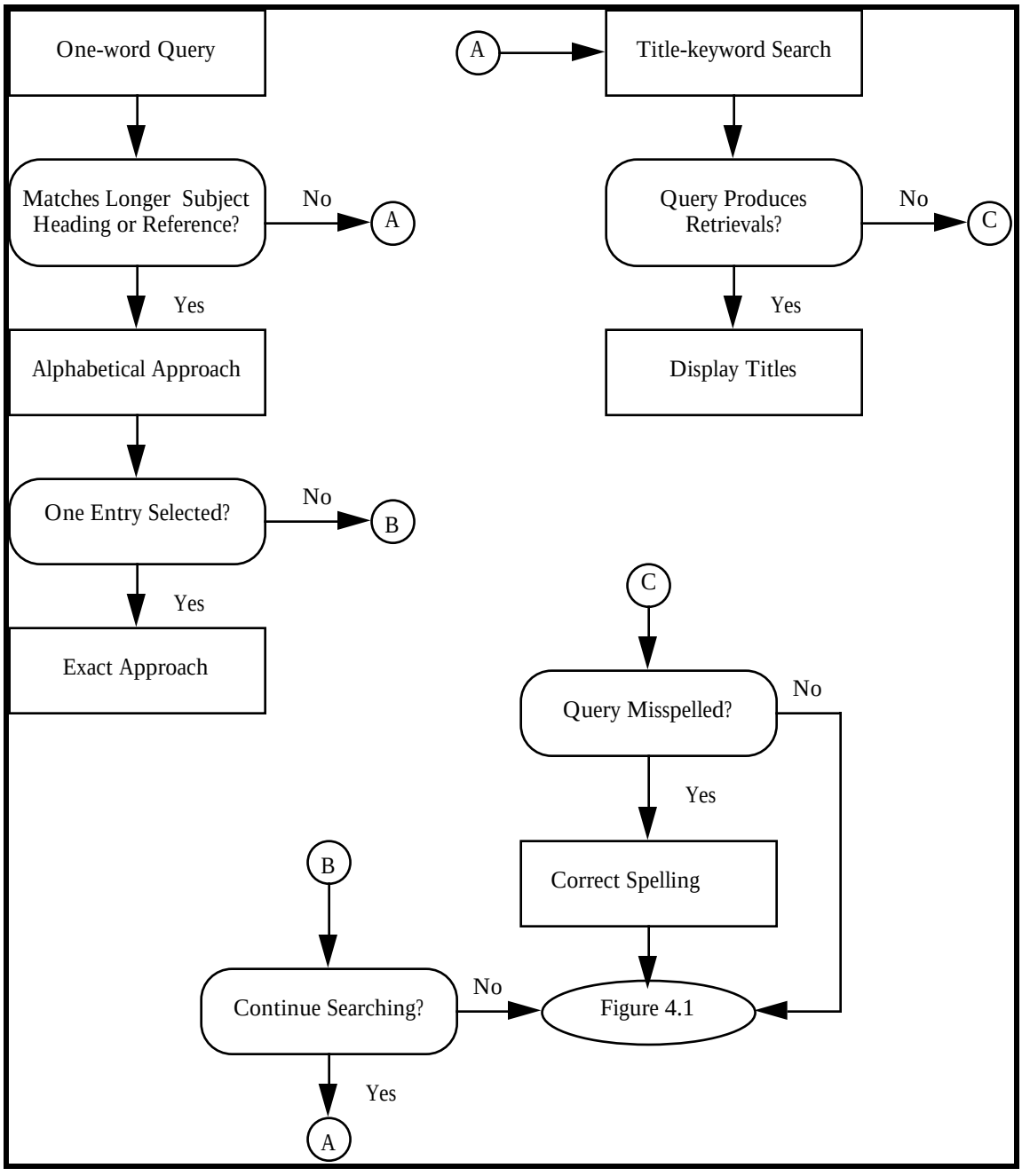

Figure 4.3. Search tree for one-word queries

The search tree for one-word queries is given in figure 4.3. One-word queries that matched the initial characters in longer controlled subject headings or *see* references were submitted to the alphabetical approach. The user's selection of a listed controlled vocabulary term invoked the exact approach. Remaining one-word queries were submitted to title-keyword searches. When title-keyword searches failed to produce

retrievals, the Blue System presented them to users and asked them to check and correct spelling. If users corrected spelling, the Blue System submitted revised queries to the initial search tree (figure 4.1) because revised queries could have matched exact or normalized forms of controlled vocabulary terms.

Remaining queries for subjects generally were composed of two or more words. Some queries matched the initial words of longer controlled vocabulary terms. These queries were submitted to the alphabetical approach (figure 4.4). The rest were submitted to a series of keyword searches that began with the keyword-in-main-heading search (figures 4.5A–4.5B). When the particular keyword search produced zero or too few retrievals, the Blue System continued searching using the next keyword approach in the series.

Figure 4.4 is a search tree for queries composed of more than one word. The Blue System responded with the alphabetical approach to queries that matched longer controlled vocabulary terms. To find additional material, the system continued searching using the keyword-in-main-heading search. This search tree took into account that some queries were partial matches of controlled vocabulary terms. For example, the query “civil rights movement” matched the first two words and part of the third word in the subject heading “Civil rights movements.” The initial system response to such queries was the alphabetical approach. When users responded to system prompts to continue searching using the original query, the Blue System continued with the results of title-keyword searches.

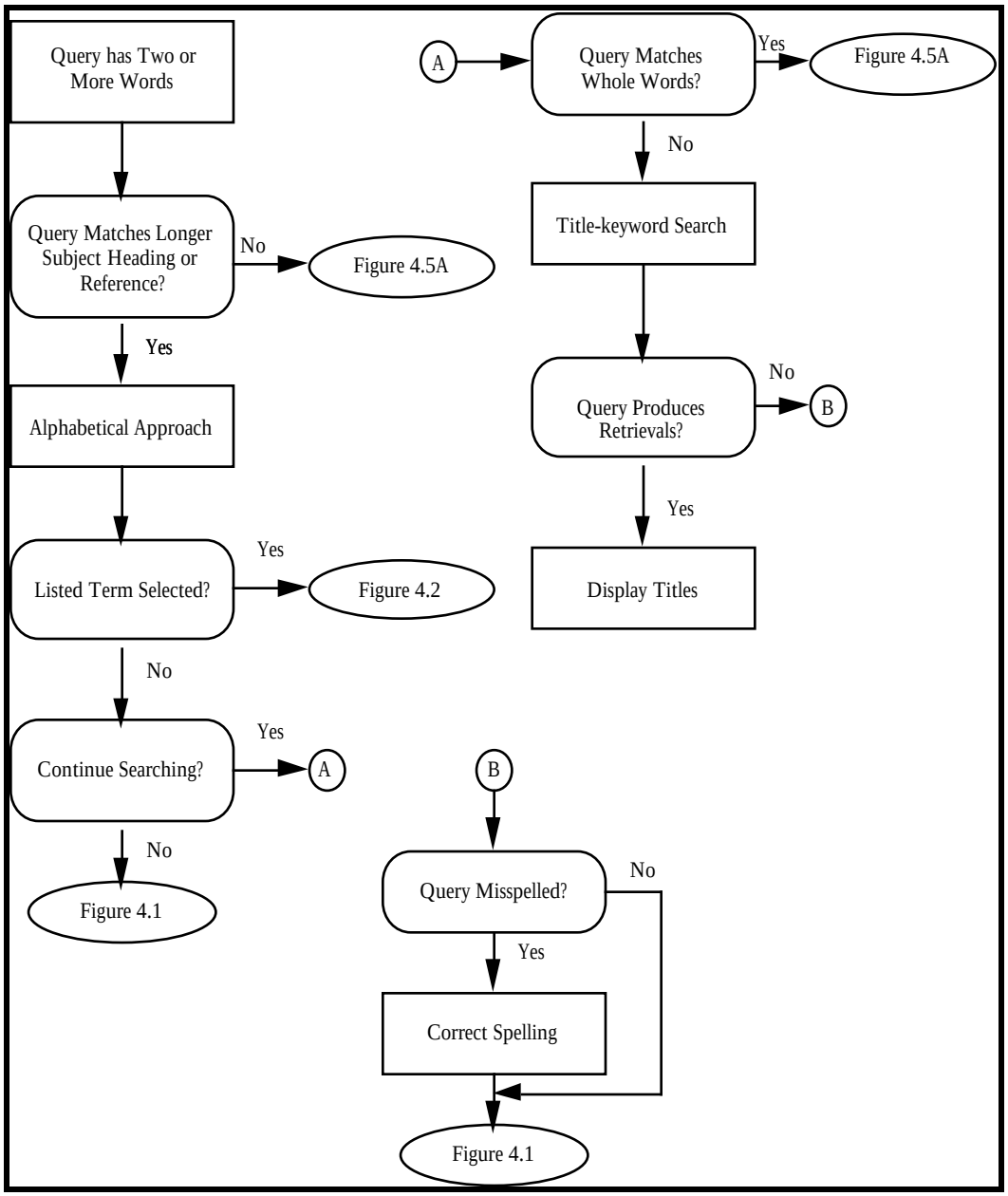

Figure 4.4. Search tree for multi-word queries featuring the alphabetical approach

Remaining queries composed of two or more words were submitted to a search tree for keyword approaches (figures 4.5A–4.5B). The search tree shown in figure 4.5A featured the submission of queries to controlled vocabulary searches.

Karen M. Drabenstott and Marjorie S. Weller

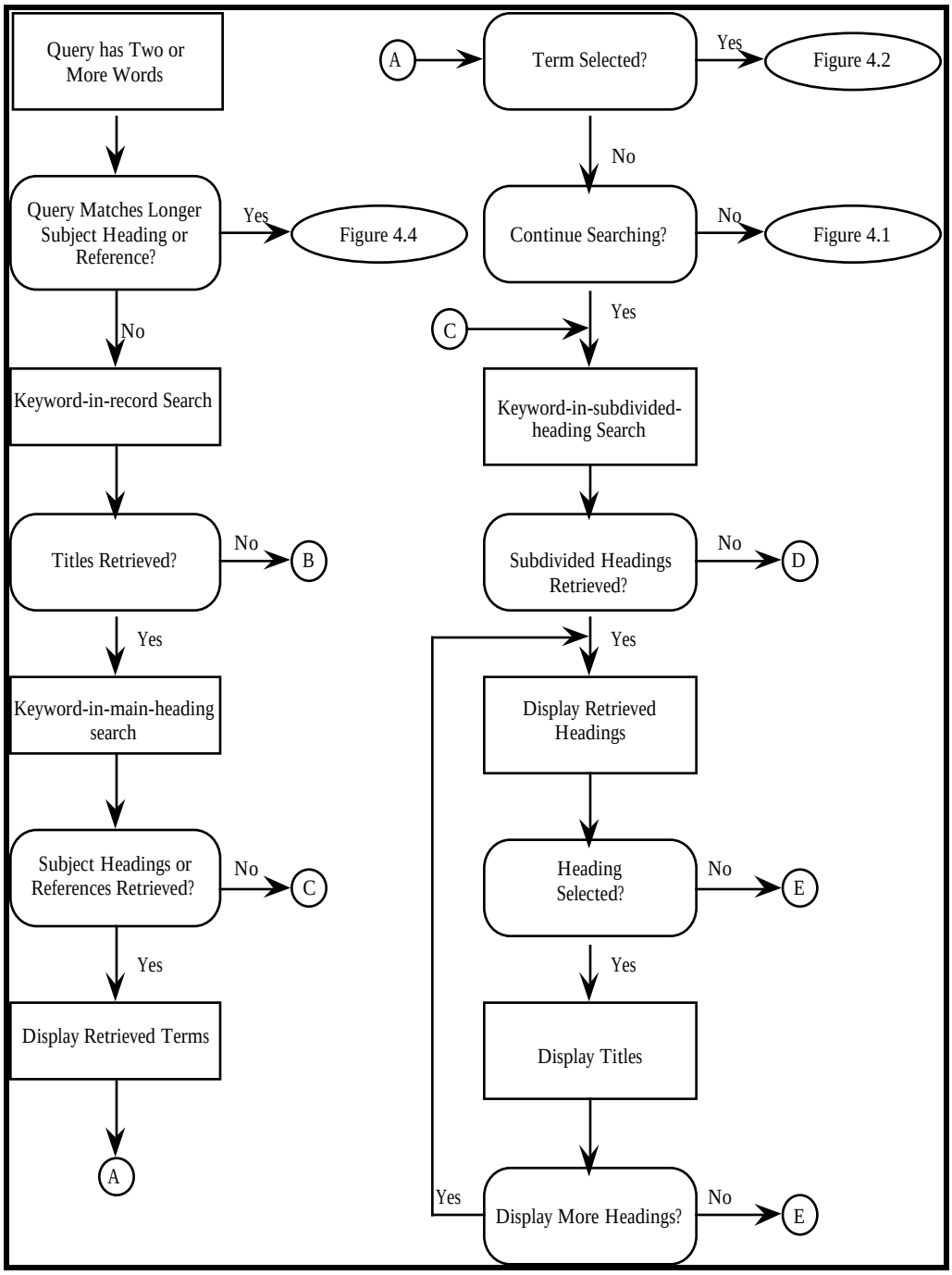

Figure 4.5A. Search tree for multi-word queries featuring keyword approaches

Prior to controlled vocabulary searches, the Blue System first used the keyword-in-record search to ensure that individual query words were posted in the catalog. If one or more query words failed to produce retrievals, the entire series of keyword searches

would fail. The Blue System asked users to check the spelling of their queries and correct them if necessary. If users made changes, the Blue System started at the beginning of the subject searching process, i.e., looking for matches of exact and normalized forms of controlled vocabulary terms. If changes were not made, this system asked users for a different query because their original one failed to produce retrievals using any keyword approach. Retrievals produced through the initial keyword-in-record search were not shown to users. Instead, the Blue System continued searching beginning with the keyword-in-main-heading search. The initial keyword-in-record search was meant to save the system additional steps searching for a query that did not retrieve any bibliographic records through the entire series of keyword searches.

Figure 4.5B features free-text searches for multi-word queries. The first free-text search was the title-keyword search. This search was followed by keyword in subject heading fields and keyword-in-record searches.

Karen M. Drabenstott and Marjorie S. Weller

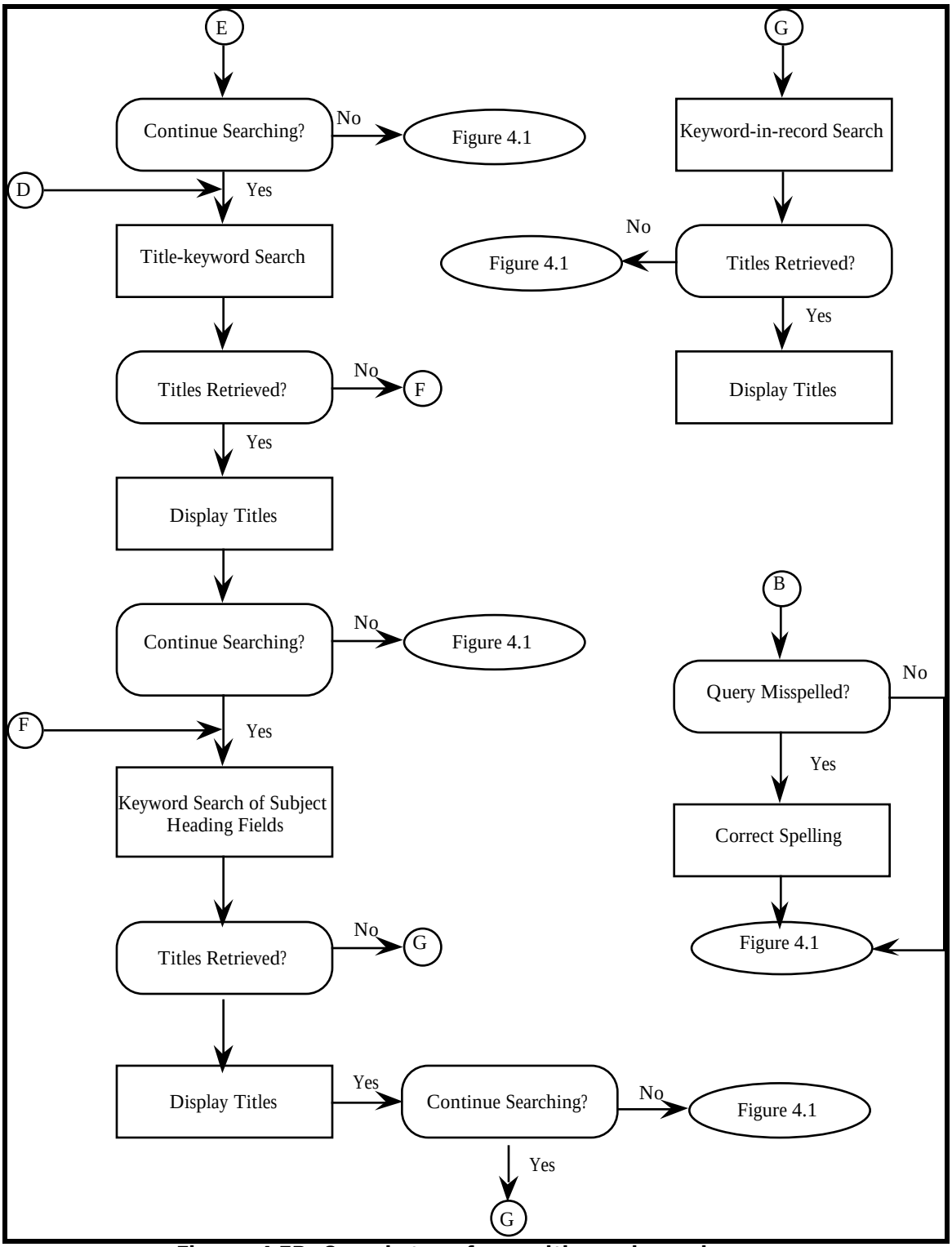

Figure 4.5B. Search tree for multi-word queries featuring keyword approaches (contd.)

4.5 Search Tree for Personal-name Queries

Both the Pinstripe and Blue Systems prompted users for surname, given name, and topical elements of their personal-name queries. When searching, the Blue System was selective about the personal-name elements it submitted to various subject searching approaches for personal names. In contrast, the Pinstripe System disregarded its knowledge of the individual elements of personal-name queries and performed keyword-in-record searches for all elements of user queries.

Table 4.1 lists the sequence of Blue-system subject searching approaches to which elements of personal-name queries were submitted. If an approach failed to produce retrievals, the Blue System chose the next approach and element(s) on the list.

Table 4.1. Sequence of personal-name query elements

	Available elements			
Approach	**Last name**	**First name**	**Middle name**	**Topic**
Queries with topics				
Keyword-in-subdivided-heading	X	X		X
Keyword-in-subdivided-heading	X			X
Keyword-in-record	X	X		X
Keyword-in-record	X			X
Queries without or omitting topic elements				
Alphabetical	X	X	X	
Alphabetical	X	X		
Alphabetical	X			

The first step was for the Blue System to single out queries containing topics. These queries were submitted to the keyword-in-subdivided-heading search followed by the keyword-in-record search in the hopes of finding headings and bibliographic records containing both name and topic elements. If the Blue System failed to find both elements, it omitted the topic element from the query and continued searching through the alphabetical approach. Personal-name queries consisting exclusively of name elements were submitted to the alphabetical approach only.

Karen M. Drabenstott and Marjorie S. Weller

4.6Chapter Summary

The search trees that governed the Blue System's selection of subject searching approaches emanated from the findings of an empirical study of user queries.

User queries for subjects generally were controlled by five search trees. The initial search tree dispatched user queries to other search trees that favored certain subject searching approaches based on summary characteristics of user queries (figure 4.1). The search tree for matches of exact and normalized forms of controlled vocabulary terms favored the exact and alphabetical approaches (figure 4.2). One-word queries were given to a search tree that favored alphabetical or title-keyword searches (figure 4.3). Queries composed of two or more words matching longer controlled vocabulary terms were submitted to alphabetical and keyword-in-heading searches (figure 4.4). Remaining queries were controlled by a search tree that submitted them to a series of keyword searches beginning with the keyword-in-record search to detect spelling errors in query words (figures 4.5A and 4.5B).

User queries for personal names were controlled by a single search tree (Table 4.1). Personal-name queries consisting exclusively of name elements were submitted to the alphabetical approach. Queries bearing topic elements were submitted to various keyword searches in the hopes of satisfying both topic and name elements. Failure to produce retrievals resulted in the alphabetical approach.

Search trees preferred two-step subject searching approaches that enlisted the catalog's controlled vocabulary, i.e., exact, alphabetical, and keyword-in-heading searches. If these searches supplied users with appropriate controlled vocabulary terms for expressing their topics of interest, users could refine their searches using related terms, subdivided forms of the matched or selected heading, or controlled vocabulary terms in close alphabetical proximity to the subject queries users enter into online catalogs.

References

Drabenstott, Karen M., and Diane Vizine-Goetz. 1994. *Using subject headings for online retrieval: Theory, practice, and potential*. San Diego: Academic Press.

Drabenstott, Karen Markey, and Diane Vizine-Goetz. 1990. "Search trees for subject searching in online catalogs." *Library Hi Tech* 8, 3: 7–20.

Mitev, Nathalie, Gillian Venner, and Stephen Walker. 1985. *Designing an online public access catalog*. London: British Lib. Library & Information Research Report 39.

O'Neill, Edward T., and Rao Aluri. 1979. *Subject heading patterns in OCLC monographic records*. Columbus, OH: OCLC. ERIC ED 183167.

Walker, Stephen, and Richard M. Jones. 1987. *Improving subject retrieval in online catalogues; 1. Stemming, automatic spelling correction and cross-reference tables*. London: British Library. British Library Research Paper no. 24.