Enhancing a New Design for Subject Access to Online Catalogs

by
Karen M. Drabenstott
with the research assistance of
Celeste M. Burman and Marjorie S. Weller

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

November 1994

@ Karen M. Drabenstott

Table of Contents

Table of Contents i

List of Tables iv

List of Figures vi

Acknowledgments vii

Obtaining copies of this report viii

1 Project Background, Objectives, and Research Questions 1
1.1 Project Overview 1
1.2 Search Tree Development 2
1.3 Using Transaction Log Analysis to Study User Queries 2
1.4 Research Questions and Methods 4
1.5 Significance of this Research Project 6
1.6 References 7

2 Overview of User Queries 10
2.1 Introduction 10
2.2 Methods Used to Manually Analyze Subject Queries 10
2.3 Subject Searching in the Four Selected Online Catalogs 13
2.3.1 Introduction 13
2.3.2 SULIRS 14
2.3.3 ORION 15
2.3.4 LS/2000 16
2.3.5 MIRLYN 16
2.4 Subject Searches and Initial Access Points 17
2.5 Initial Access Points Discarded from the Analyses 20
2.5.1 Introduction 20
2.5.2 Known-item Access Points 22
2.5.3 Other Discarded Initial Access Points 23
2.6 Chapter Summary 25
2.7 References 26

3 Search Trees for Subject Searching 27
3.1 Introduction 27
3.2 Search Tree Design 27
3.3 Initial Search Tree 28
3.4 Search Trees for Subject Queries Generally 29
3.5 Search Tree for Personal-name Queries 39

3.6 Chapter Summary 40
3.7 References 40

4 Search-tree Selections for End-user Queries 41
4.1 Introduction 41
4.2 Facts on Queries for Subjects Generally 41
4.2.1 Topical Subjects 41
4.2.2 Geographic Names 42
4.2.3 Combinations of Topical Subjects and Geographic Names 43
4.2.4 Corporate Names 44
4.3 Queries for the Exact Approach 45
4.3.1 Introduction 45
4.3.2 Types of Exact Matches 46
4.3.3 Match Satisfaction 48
4.4 Queries for the Alphabetical Approach 50
4.4.1 Introduction 50
4.4.2 Types of Alphabetical Matches 51
4.4.3 Match Satisfaction 52
4.5 Queries for Keyword-in-heading Searches 56
4.5.1 Introduction 56
4.5.2 Types of Keyword-in-heading Matches 57
4.5.3 Match Satisfaction 57
4.6 Queries for Keyword Searches 61
4.6.1 Introduction 61
4.6.2 Types of Keyword Matches 62
4.6.3 Match Satisfaction 62
4.7 Chapter Summary 65

5 No Matches 68
5.1 Introduction 68
5.2 Basic Information 68
5.3 Redesigned Search Trees 70
5.3.1 Initial Search Tree 70
5.3.2 Search Tree for One-word Queries 71
5.3.3 Search Trees for Multi-word Queries 74
5.4 Testing Enhanced Search-tree Configuration 81
5.4.1 One-word Queries 81
5.4.2 Two-word Queries 81
5.4.3 Queries Exceeding Two Words 83
5.5 Chapter Summary 86
5.6 References 87

6 Queries for Personal Names 88
6.1 Introduction 88
6.2 Combinations of Topical and Personal-Name Elements 89
6.2.1 Introduction 89
6.2.2 Keyword and Alphabetical Matches 90
6.3 Alphabetical Matches 92
6.3.1 Introduction 92
6.3.2 Form of Name Entered 93
6.3.3 Alphabetical Approach for Personal-name Queries 96
6.3.4 Match Satisfaction 97

6.4 Chapter Summary 99
6.5 References 101

7 Classification for Subject Searching 102
7.1 Introduction 102
7.2 Methodology 103
7.3 Summarization Using DDC 104
7.3.1 Acid Rain 104
7.3.2 Costa Rica 107
7.3.3 Greek Sculpture 110
7.3.4 Pornography 111
7.3.5 Racism 114
7.4 Summarization Using LCC 116
7.4.1 Acid Rain 116
7.4.2 Costa Rica 119
7.4.3 Greek Sculpture 122
7.4.4 Pornography 125
7.4.5 Racism 127
7.5 Using Classification to Summarize Large Retrievals 131
7.6 Using Classification to Increase Very Low Retrievals 133
7.7 Chapter Summary 136
7.8 References 138

8 Highlights of Project Activities and Findings 140
8.1 Project Overview 140
8.2 Research Questions and Methods 140
8.3 Subject Searches in Systems Governed by Search Trees 141
8.4 Enhancements to Search Trees 143
8.5 Handling Too Many Retrievals 144
8.6 Handling Too Few Retrievals 145
8.7 Improving Long Lists of Retrieved Titles 146
8.8 References 146

Appendixes

A DDC Retrievals for “acid rain” 147
B DDC Retrievals for “costa rica” 148
C DDC Retrievals for “greek sculpture” 150
D DDC Retrievals for “pornography” 151
E DDC Retrievals for “racism” 152
F LCC Retrievals for “acid rain” 154
G LCC Retrievals for “costa rica” 155
H LCC Retrievals for “greek sculpture” 157
I LCC Retrievals for “pornography” 159
J LCC Retrievals for “racism” 160

List of Tables

2.1 Transaction Log Data 13
2.2 Types of Initial Access Points 17
2.3 Fact Sheet for Initial Access Points in Subject Searches 18
2.4 Fact Sheet for Discarded Initial Access Points 21
3.1 Sequence of personal-name query elements 39
4.1 Fact Sheet for Queries for Topical Subjects 41
4.2 Fact Sheet for Subject Queries for Geographic Names 42
4.3 Fact Sheet for Queries for Topical Subjects and Geographic Names 43
4.4 Fact Sheet for Subject Queries for Corporate Names 44
4.5 Fact Sheet for Queries for the Exact Approach 45
4.6 Types of Exact Matches 46
4.7 Normalized Matches 47
4.8 Satisfaction Categories for Exact Matches 48
4.9 Off-the-mark Matching Terms 48
4.10 Queries Too Broad or Narrow to Express Users' Topics of Interest 49
4.11 Fact Sheet for Queries for the Alphabetical Approach 50
4.12 Types of Partial Matches 51
4.13 Partial Matches for the Alphabetical Approach 52
4.14 Browsing for Suitable Terms 54
4.15 "Off-the-mark" and "Too Narrow" Categories (Partial Matches) 54
4.16 Fact Sheet for Queries for Keyword-in-heading Searches 56
4.17 Types of Keyword-in-heading Matches 57
4.18 "Off-the-mark" and "Too Narrow" Categories (Keyword-in-heading Matches) 59
4.19 Fact Sheet for Queries for Keyword Searches 61
4.20 Types of Keyword Matches 62
4.21 Strategies for Enhancing "Too Few Retrievals" 63
4.22 "Off-the-mark" and "Too Narrow" Categories (Keyword Matches) 65
5.1 Fact Sheet for Queries for No Matches 68
5.2 Matches of Two-word Queries 81
5.3 Matches of Two-word Queries in Subject- or Form-specific Databases 82
5.4 Alphabetical Matches of Two-word Queries 83
5.5 Matches of Queries Exceeding Two Words 84
5.6 Matches of Queries Exceeding Two Words in Subject- or Form-specific Databases 85
5.7 Retrievals from Best-match Searches 85
6.1 Fact Sheet for Personal-name Queries Bearing Topical and Name Elements 89
6.2 Types of Keyword Matches 90
6.3 Fact Sheet for Personal-name Queries Bearing Name Elements Only 92

6.4 Types of Alphabetical Matches 96
7.1 DDC Retrievals for “acid rain” 104
7.2 First-level DDC Captions for “acid rain” 105
7.3 Second-level DDC Captions for “acid rain” in the Social Sciences 105
7.4 EDDC Retrievals for “acid rain” 106
7.5 DDC Retrievals for “costa rica” 107
7.6 First-level DDC Captions for “costa rica” 107
7.7 Second-level DDC Captions for “costa rica” in the Social Sciences 108
7.8 Second-level DDC Captions for “costa rica” in Geography and History 108
7.9 EDDC Retrievals for “costa rica” 109
7.10 DDC Retrievals for “greek sculpture” 110
7.11 First-level DDC Captions for “greek sculpture” 110
7.12 Second-level DDC Captions for “greek sculpture” 111
7.13 DDC Retrievals for “pornography” 111
7.14 First-level DDC Captions for “pornography” 112
7.15 Second-level DDC Captions for “pornography” 112
7.16 EDDC Retrievals for “pornography” 113
7.17 DDC Retrievals for “racism” 114
7.18 First-level DDC Captions for “racism” 114
7.19 Second-level DDC Captions for “racism” 115
7.20 EDDC Retrievals for “racism” 116
7.21 LCC Retrievals for “acid rain” 116
7.22 First-level LCC Captions for “acid rain” 117
7.23 Second-level LCC Captions for “acid rain” in Technology 118
7.24 Third-level LCC Captions for “acid rain” in Technology 118
7.25 LCC Schedule Retrievals for “acid rain” 119
7.26 LCC Retrievals for “costa rica” 119
7.27 First-level LCC Captions for “costa rica” 120
7.28 Second-level LCC Captions for “costa rica” in History 120
7.29 Second-level LCC Captions for “costa rica” in the Social Sciences 121
7.30 LCC Schedule Retrievals for “costa rica” 122
7.31 LCC Retrievals for “greek sculpture” 122
7.32 First-level LCC Captions for “greek sculpture” 123
7.33 Second-level LCC Captions for “greek sculpture” in the Fine Arts 124
7.34 LCC Schedule Retrievals for “greek sculpture” 124
7.35 LCC Retrievals for “pornography” 125
7.36 First-level LCC Captions for “pornography” 125
7.37 Second-level LCC Captions for “pornography” in the Social Sciences 126
7.38 LCC Schedule Retrievals for “pornography” 126
7.39 LCC Retrievals for “racism” 127
7.40 First-level LCC Captions for “racism” 128
7.41 Second-level LCC Captions for “racism” in World History 128
7.42 Second-level LCC Captions for “racism” in the History of North and South America 129
7.43 Second-level LCC Captions for “racism” in the Social Sciences 129
7.44 LCC Schedule Retrievals for “racism” 130
7.45 Using DDC Numbers to Increase Retrievals 134
7.46 Using LCC Numbers to Increase Retrievals 134
8.1 Matches of User Queries and Subject Searches 142

List of Figures

1.1 Schedule of major project activities 6
2.1 Portion of a SULIRS Log 12
2.2 Types of initial access points 18
2.3 Discarded initial access points 21
3.1 Initial search tree 28
3.2A Search tree for the exact approach 30
3.2B Search tree for the exact approach (contd.) 31
3.3 Search tree for one-word queries 33
3.4 Search tree for multi-word queries featuring the alphabetical approach 35
3.5A Search tree for multi-word queries featuring keyword approaches 36
3.5B Search tree for multi-word queries featuring keyword approaches (contd.) 38
4.1 Satisfaction categories for partial matches 53
4.2 Satisfaction categories for keyword-in-heading matches 58
4.3 Satisfaction categories for keyword matches 63
5.1 Elements in non-matching queries 69
5.2 Redesigned initial search tree 71
5.3A Redesigned search tree for one-word queries 72
5.3B Redesigned search tree for one-word queries (contd.) 74
5.4A Redesigned search tree for multi-word queries 76
5.4B Redesigned search tree for multi-word queries (contd.) 78
5.4C Redesigned search tree for multi-word queries (contd.) 80
6.1 Forms of personal-name queries 94
6.2 Satisfaction categories for alphabetical matches 98
8.1 Matches of user queries and subject searches 141

Acknowledgments

The Michigan project team consisted of Karen Markey Drabenstott, Celeste M. Burman, and Marjorie S. Weller.

In this report, we describe the enhancement of a new design for subject access to online catalogs. The OCLC Online Computer Library Center, Inc., provided support for the study through its Library and Information Science Research Grant Program. Dr. Edward T. O'Neill served as the OCLC Project Monitor.

We are especially grateful to Beth Warner, Librarian and Head, Systems Office, University of Michigan Library, who supplied the project team with transaction log records from the Library's MIRLYN online catalog. We also thank Phyllis A. Valentine, Senior Associate Librarian, Systems Office, who generated MIRLYN transaction log data from our specifications.

The project team provided the principal author with invaluable assistance throughout the project. Referring to the categorizations of end-user queries that Martha O'Hara Conway completed in a previous, related project, Celeste Burman applied a similar categorization scheme with different objectives to analyze a few thousand end-user queries. She also manually entered thousands of outline records from the Library of Congress Classification into a FoxPro database. Marjorie Weller programmed the FoxPro database for subject searching. Both Celeste and Marjorie completed project tasks after they left the School of Information and Library Studies to assume new jobs at the Detroit Public Library and University of Michigan School of Medicine, respectively. The principal author is grateful to them for their dedication and contributions to the project.

The end-user queries that were categorized in this project were collected and analyzed in an earlier, related study. The authors thank University Librarians David M. Stam at Syracuse University, Paul Willis at the University of Kentucky, and Gloria Werner at the University of California, Los Angeles, who consented to our analysis of transaction logs of their respective libraries' online catalogs. We also thank Donald Riggs who encouraged us to include queries from the University of Michigan's MIRLYN online catalog in this study.

Several colleauges provided insight and were themselves sources of information, particularly Terry Ferl, Carolyn O. Frost, Diane Vizine-Goetz, and Amy J. Warner.

Obtaining copies of this report

Enhancing a New Design for Subject Access to Online Catalogs may be obtained over the Internet using anonymous FTP. The FTP host name is "sils.umich.edu." When the system asks you for a user name and password, type the user name "anonymous" and type your electronic mail address (or last name) for the password. The directory containing this report is /pub/papers/ENHANCE.

Although the files in the anonymous FTP directory are made available in several formats, we strongly suggest you obtain this report in Postscript or PDF (Adobe Acrobat) file formats to preserve layout, paging, figures, tables, and type fonts. The formats of files in the anonymous FTP directory are as follows:

FILE NAME:	DESCRIPTION AND FORMAT:
ENHANCE.pdf	Report in PDF (Adobe Acrobat) format
ENHANCE.ps	Report in Postscript format
ENHANCE.sit.hqx	Report in Microsoft Word 5.1a (Mac, stuffed and binhexed)

To retrieve the ENHANCE.sit.hqx file, you will need a copy of the BinHex and UnStuffIt utilities for Macintosh. These are free, public domain utilities which can be obtained via anonymous FTP from "archive.umich.edu" in the directory /mac/utilities/compression. To use the file, you will need Microsoft Word 5.1a for the Macintosh.

Published (bound and printed) copies of *Enhancing a New Design* may be obtained from the School of Information and Library Studies at the following rates: (1) $10 prepaid, 4th class delivery in the United States, (2) $12 prepaid, first class delivery in the United States, (3) $15 prepaid, first class delivery in Canada, or (4) $25 prepaid, air mail delivery outside the United States and Canada. Please make checks payable to the University of Michigan and send them to the School of Information and Library Studies, University of Michigan, 304 West Engineering Building, 550 East University Avenue, Ann Arbor, Michigan 48109–1092 USA. The authors have submitted the report to ERIC, and, pending acceptance and processing, will be available in microfiche and hardcopy formats from ERIC.

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.

Celeste M. Burman categorized the thousands of user-entered queries discussed in this research project, manually coded and entered them into a microcomputer-based statistical analysis program. She also manually entered the outline headings of the Library of Congress Classification (LCC) schedules into a searchable FoxPro database. Celeste is currently a Reference Librarian II at the downtown branch of the Detroit Public Library. She received her B.A. and M.I.L.S. from the University of Michigan.

Marjorie S. Weller designed, developed, and implemented a searchable database of LCC outline headings 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.

1 Project Background, Objectives, and Research Questions

1.1Project Overview

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 are the number of words in user queries, the extent to which user queries match controlled vocabulary terms, and the ability of user queries to produce retrievals in response to certain subject searching approaches.

A new design for subject access to online catalogs enlists search trees to identify these characteristics, control system responses, and determine appropriate subject searching approaches to user queries. Search trees are a set of paths with branches or choices that enable systems to carry out the most sensible search approach at each stage of the search. The search trees that are the focus of this research project were developed from an empirical study of subject queries users enter into online catalogs (Drabenstott and Vizine-Goetz 1990, 1994; Vizine-Goetz and Drabenstott 1991). Search trees were limited to subject searching approaches implemented in *operational* online catalogs, i.e., exact, alphabetical, and various keyword approaches. These approaches, however, failed to produce retrievals for some queries.

The purpose of this research project was to enhance the search trees with new subject searching approaches to enable online catalogs to respond with useful information for the most difficult user queries. To achieve this goal, the researchers extracted subject queries from the online catalog transaction logs of four research libraries and determined the particular subject searching approach that search trees would invoke. The researchers paid special attention to failed queries, that is, queries that failed to produce retrievals, and developed subcategories of these queries corresponding to the subject searching approaches that would provide useful retrievals. They scrutinized subcategories of subject queries to determine characteristics of queries that online systems could recognize without the aid of human intermediaries. They then developed specifications for subject searching approaches that would be likely to provide useful information for subcategorized subject queries. Subject searching approaches were variations of approaches available in operational online catalogs.

They were also entirely new approaches, ones drawn from information retrieval research, and ones that enlisted library classifications.

1.2Search Tree Development

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 title-keyword searches (Hildreth 1989b, 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.

Underlying the original definition for search trees was a much different objective. Search trees selected the subject searching approach most likely to produce useful information in response to user queries. The search trees that were the focus of this study were derived from findings of an empirical study of user queries (Drabenstott and Vizine-Goetz 1990; 1994). These search trees emphasized subject searching approaches using 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) system (O’Neill and Aluri 1979, 5). They favored approaches that enlisted the catalog’s controlled vocabulary because such approaches exemplified the strategies expert searchers pursued.

1.3Using Transaction Log Analysis to Study User Queries

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. 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.

Transaction log analysis also has disadvantages (Hildreth 1985; Hancock et al. 1990; Kurth 1994). Two drawbacks pertinent to the proposed study are the researcher's inability to know exactly what users are looking for and when users begin and end their searches. Information regarding the former can be obtained by supplementing transaction logs with observations of user searches, personal interviews, or a combination of the two. The latter can be overcome by manual analyses of transaction logs. They are more accurate than computer analyses because human intermediaries can enlist changes in both time stamps and the meaning of user queries to demarcate individual searches.

Transaction log analysis has been used to study user-entered access points since the time of the Council on Library Resources-supported nationwide survey of online catalogs (Borgman 1983; Larson 1983; Tolle 1983). The objective of the many researchers who have studied transaction logs has been to recommend improvements to the subject searching capabilities of online catalogs. Lester (1989, 58–98) gives examples of recommended improvements:

- 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.

Such lists of improvements are of limited use for the following reasons: (1) they do not tell us which improvement will improve the prospects for success for the greatest number of queries entered into online catalogs, (2) they do not tell us whether there are certain characteristics of user queries that make them better suited for particular subject searching improvements, and (3) they do not tell us which subject searching capabilities in our existing online catalogs are currently working satisfactorily on user queries and the characteristics of these queries. The closest that any study comes to shedding light on these three points is a recent empirical study of the subject terms users entered into online catalogs that demonstrated "the retrieval processes of right truncation, string searching, and keyword searching each make significant improvements in match success with the Library of Congress subject headings" (Lester 1989, 267).

The principal investigator of the proposed study was one of two principals who recognized the need for a study to determine which retrieval processes were best suited to the wide variety of user queries. Supported by the Council on Library Resources, the two principals conducted an empirical study of the subject terms users entered into online catalogs and developed search trees based on the results (Drabenstott and Vizine-Goetz 1990; 1994). Search trees controlled system responses and determined appropriate subject searching approaches to user queries. They placed the burden of determining which approach was likely to produce useful information on the system.

The two researchers acknowledged the inability of subject searching approaches in operational online catalogs to produce retrievals for some subject queries. They described several techniques used in information retrieval research — stemming, automatic spelling correction, statistical retrieval methods — that might be helpful for producing retrievals and called for additional research to determine the most effective techniques to apply to such queries (Drabenstott and Vizine-Goetz 1994, 290–3).

The time was right for a new empirical study of user queries that focused on subject queries for which online systems were unable to produce retrievals. This report describes the findings of such a study. It recommends entirely new retrieval methods and methods from information retrieval research that would be particularly suited to cataloging databases in which subject description is limited to title fields and subject headings based on the LCSH system. It also results in an enhancement of search trees to allow for newly recommended approaches.

1.4Research Questions and Methods

The objective of this research project is to enhance the existing configuration of search trees with new subject searching approaches that are not available in operational online catalogs to provide useful information for the most difficult user queries. The study answers five research questions:

1. To what extent do user queries fail to produce retrievals through the subject searching approaches in the existing search-tree configuration?
2. What subject searching approaches would provide useful retrievals for these failed queries?
3. To what extent do user queries match controlled vocabulary terms that are not posted in the online catalog searched?
4. What subject searching approaches would provide useful retrievals for user queries that match unposted controlled vocabulary terms?
5. What enhancements are needed to the existing search-tree configuration to improve the quality and responsiveness of online catalogs to the user queries selected for study in this project?

In a previous research project sponsored by the Council on Library Resources, the Michigan project team performed a manual analysis of over fifteen hundred subject queries from the SULIRS, ORION, and LS/2000 online catalogs at Syracuse University, University of California, Los Angeles (UCLA), and University of Kentucky, respectively. This analysis resulted in the development of the original configuration of search trees (Drabenstott and Vizine-Goetz 1990; 1994). In this research project, the project team analyzed the same set of queries and analyzed over four hundred additional queries from the transaction logs of MIRLYN, the online catalog of the University of Michigan. The objective of the latter analysis was to answer the five research questions listed above about enhancing the existing configuration of search trees. The benefits of using queries from the previous study in this study were: (1) the query selection process was quicker than working from scratch because queries that were candidates for inclusion in the this study were already coded in certain categories and could be selected from an existing data file using SYSTAT, a microcomputer-based statistical analysis program, (2) the worksheets for selection of MIRLYN queries could be based on categories in worksheets from the previous study, and (3) the amount of time spent training staff would be reduced because some procedures and coding tools developed for the previous study could be applied for the proposed study.

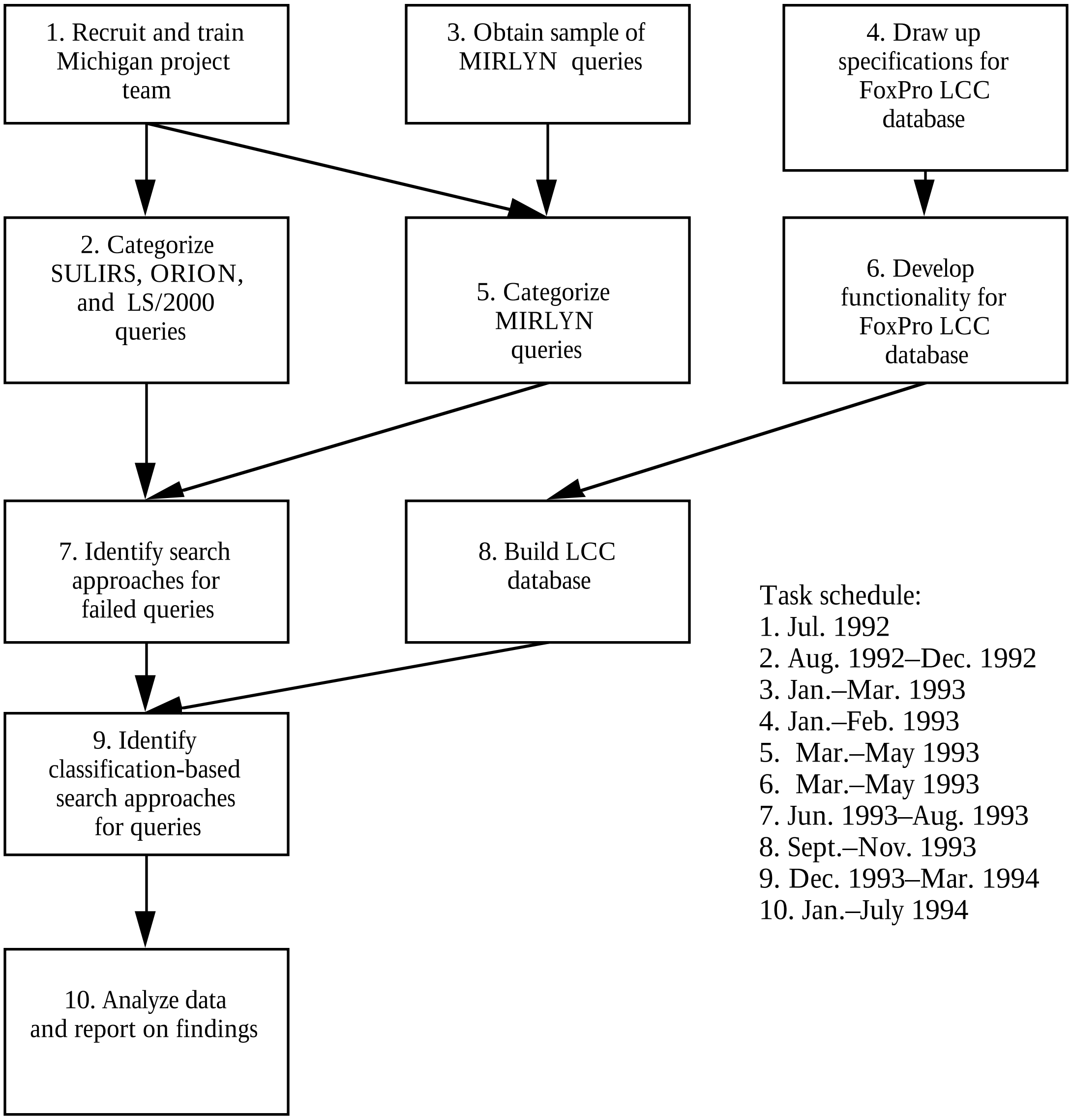
The Michigan project team selected the initial queries users entered in subject searches from the four libraries' transaction logs. The team categorized queries by the type(s) of elements present in them: (a) topical subjects, (b) corporate names, (c) geographic names, (d) personal names, and (e) combinations of two or more elements (a–d).

The team developed subcategories of selected queries corresponding to the search-tree subject searching approaches that would provide useful retrievals. Staff also scrutinized subcategories of subject queries to determine whether they had certain characteristics that online systems could recognize without the help of human intermediaries. Subject searching enhancements to search trees capitalized on the ability of systems to automatically recognize queries with these certain characteristics and passed such queries to newly-defined subject searching approaches for which they were suited.

The Michigan project team used SYSTAT to derive the summary statistics necessary for answering the project's first and third research questions regarding the extent to which user queries fail to produce retrievals. Staff were guided in determining the answers to the project's second and fourth research questions regarding the subject searching approaches that would provide useful retrievals to failed queries by their knowledge of subject searching capabilities in experimental online catalogs and other information retrieval systems that are described in the professional literature. To simulate the performance of these approaches, staff conducted online searches for certain queries in the operational and experimental online catalogs available through the Internet. They also gauged the performance of the Dewey Decimal Classification (DDC) and Library of Congress Classification (LCC) for queries that retrieved too many or too few retrievals. We used the Electronic DDC to assess the performance of the DDC. We manually entered LCC synopses and outlines into a searchable database that was developed for this project specially using the microcomputer-based FoxPro database management system and searched this database to assess the performance of LCC.

Answers to the first four research questions provided us with the insight to answer the project's fifth research question. We enhanced the search trees with new subject searching approaches that are not available in operational online catalogs to enable systems to respond with useful information for the most difficult user queries.

Figure 1.1 shows a schedule of major project activities.


1. Recruit and train Michigan project team
3. Obtain sample of MIRLYN queries
4. Draw up specifications for FoxPro LCC database
2. Categorize SULIRS, ORION, and LS/2000 queries
5. Categorize MIRLYN queries
6. Develop functionality for FoxPro LCC database
7. Identify search approaches for failed queries
8. Build LCC database
9. Identify classification-based search approaches for queries
10. Analyze data and report on findings

Task schedule:
1. Jul. 1992
2. Aug. 1992–Dec. 1992
3. Jan.–Mar. 1993
4. Jan.–Feb. 1993
5. Mar.–May 1993
6. Mar.–May 1993
7. Jun. 1993–Aug. 1993
8. Sept.–Nov. 1993
9. Dec. 1993–Mar. 1994
10. Jan.–July 1994

Figure 1.1. Schedule of major project activities

1.5Significance of this Research Project

The objective of this research project responds to the key challenge facing today's online catalog system designers, viz. "to exploit the science and technology of automated information retrieval to achieve the 'best' retrieval for a given user query in an inherently imprecise and uncertain situation" (Hildreth 1989b, 46). This research project builds on the existing configuration of search trees that enlists exact, alphabetical, and keyword subject searching approaches in response to user queries. Such approaches characterize first-and second-generation online catalogs (Hildreth 1984).

Findings of information retrieval research have characterized the search functionality of third-generation online catalogs (Hildreth 1984). Such functionality, e.g., extended Boolean processing, fuzzy-set retrieval, statistical retrieval methods, is virtually non-existent in operational online catalogs (Walker 1989, 103; Hildreth 1989a, 3). It has been limited to a handful of experimental systems, e.g., OKAPI (Walker 1989), SPRILIB (Porter and Galpin 1988), and INSTRUCT (Hendry, Willett, and Wood 1986a, 1986b), and a single operational system, CITE, that was available to searchers at the National Library of Medicine in the mid 1980s (Doszkocs 1983). Reasons why third-generation functionality has not been implemented in existing online catalogs are (Mitev, Venner, and Walker 1985, 164): (1) it is difficult to reconcile with Boolean searching on which most existing systems are based, (2) it is computationally heavy, and (3) it comes into its own on queries containing three or more words but such queries are a minority of online catalog queries (Walker 1991; Van Pulis and Ludy 1988, 527; Drabenstott and Vizine-Goetz 1994, 157).

This project builds on a previous study of user queries that demonstrated the inability of subject searching approaches in first- and second-generation online catalogs to produce retrievals for certain user queries for subjects (Drabenstott and Vizine-Goetz, 1994). It identifies characteristics of user queries that cannot be handled by subject searching approaches in operational online catalogs. It considers related bibliographic files, i.e., library classifications, subject heading authority files, as sources of subject terminology and structured outlines for retrieval and browsing. It looks to information retrieval research for retrieval methods that would be effective in providing useful information for these queries and makes enhancements to the search trees to include these methods.

Online system designers have been reluctant to introduce third-generation functionality for the reasons given above. Although this research project results in enhanced search trees that include third-generation search functionality, it demonstrates that third-generation functionality is needed for a fraction of user queries for subjects. The project determines the extent to which user queries fail to produce retrievals through subject searching approaches in operational online catalogs. It identifies the characteristics of such queries, defines the retrieval methods that are likely to produce useful information, and enhances the search trees to include them. Methods could be derived from information retrieval research or be connected with subject retrieval and browsing of related bibliographic files.

1.6 References

Borgman, Christine L. 1983. *End user behavior on The Ohio State University Libraries' online catalog: a computer monitoring study*. Dublin, Ohio: OCLC. OCLC Research Report Series OCLC/OPR/RR-83/7.

Doszkocs, Tamas E. 1983. "CITE NLM: Natural-language searching in an online catalog." *Information Technology and Libraries* 2, 4: 364–380.

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

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

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

Hendry, Ian G., Peter Willett, and Frances E. Wood. 1986a. "INSTRUCT: a teaching package for experimental methods in information retrieval. Part I. The users' view." *Program* 20, 3 (July): 245–63.

Hendry, Ian G., Peter Willett, and Frances E. Wood. 1986b. "INSTRUCT: a teaching package for experimental methods in information retrieval. Part II. Computational aspects." *Program* 20, 4 (October): 382–93.

Hildreth, Charles R. 1989a. "General introduction: OPAC research: laying the groundwork for future OPAC design." In *The online catalogue: developments and directions*, edited by Charles R. Hildreth, 1–24. London: The Library Association.

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

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

Hildreth, Charles R. 1984. "Pursuing the ideal: generations of online catalogs." In *Online catalogs/online reference: Converging trends*, edited by Brian Aveney and Brett Butler, 31–56. Chicago: American Library Association.

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

Larson, Ray R. 1983. Users look at online catalogs; part 2: interacting with online catalogs; final report to the Council on Library Resources. Berkeley, Calif.: University of California. ED 231401.

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.

Mitev, Nathalie Nadia, Gillian M. Venner, and Stephen Walker. 1985. *Designing an online public access catalogue*. London: British Library. British Library Research Report 39.

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.

Porter, Martin, and Valerie Galpin. 1988. "Relevance feedback in a public access catalogue for a research library: Muscat at the Scott Polar Research Institute." *Program* 22, 1 (January): 1–20.

Tolle, John E. 1983. *Current utilization of online catalogs: transaction log analysis. Vol. 1 of final report to the Council on Library Resources.* Dublin, Ohio: OCLC. OCLC Research Report Series OCLC/OPR/RR-83/2. ED 231402.

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

Vizine-Goetz, Diane, and Karen M. Drabenstott. 1991. "Computer and manual analysis of subject terms entered by online catalog users." In *Proceedings of the 54th ASIS annual meeting*, edited by Jose-Marie Griffiths, 156–61. Medford, NJ: Learned Information.

Walker, Stephen. 1991. "Subject access in online catalogues." Unpublished notes given in a lecture in Copenhagen, 11 April.

Walker, Stephen. 1989. “The OKAPI online catalogue research projects.” In *The online catalogue: developments and directions*, edited by Charles R. Hildreth, 84–106. London: The Library Association.

2 Overview of User Queries

2.1Introduction

This chapter is an orientation to the subject terms users enter into online catalogs. It reports on the number and percentages of legitimate user queries. To determine the legitimacy of user queries, we placed queries into legitimate subject-term categories for single-element terms and multiple-element terms and into categories for terms discarded from subsequent analyses, i.e., known-item access points and "other" access points. Legitimate subject-term and known-item subcategories were named for the topical and name elements present in access points. Examples of queries in the topical-personal name subcategory are "clarence darrow's relegious [sic] views" and "hitler's speeches." Both examples had personal-name elements (i.e., "clarence darrow's" and "hitler's") and topical-subject elements (i.e., "relegious [sic] views" and "speeches"). Queries that were not legitimate were discarded from subsequent analyses.

Analyses of user queries from the online catalogs of Syracuse University, UCLA, and University of Kentucky in this and subsequent chapters resulted in statistics that were almost the same as statistics in a previous study (Drabenstott and Vizine-Goetz 1994). Although the queries from these three institutions were the same in both studies, one of the two researchers who categorized queries was different. On occasion, we disagreed with the categorization in the previous study or we had learned something new about the query that made us apply a different categorization. For example, the query "beer mathematics" was originally placed in the "playing" category because we thought that the user was entering a subject search for "mathematics," and, at the same time, daydreaming about drinking beer. In this study, we placed this query in the known-item search categorization categorization because a known-item search for this query in several online catalog databases resulted in the retrieval of titles on mathematics by Gerald Alan Beer.

2.2Methods Used to Manually Analyze Subject Queries

We randomly selected several complete days of activity from terminals accessed by library patrons at Syracuse, UCLA, and University of Kentucky. The process of selecting online catalog searching activity was somewhat more complex for the University of Michigan's

MIRLYN online catalog than at the other three libraries. When Michigan users signed onto MIRLYN, the system assigned a unique terminal number to the connected terminal. As long as users remained connected to MIRLYN and/or the system remained up and running, the MIRLYN-assigned terminal number did not change. Many users exited MIRLYN and signed onto the university's computers to read their electronic mail or perform other tasks. When they later signed onto MIRLYN, the system assigned the session a new terminal number.

To ensure that we collected *end-user* activity from MIRLYN terminals, we randomly selected morning, afternoon, and evening time periods. During selected periods, a monitor observed user activity and entered a command into MIRLYN to verify MIRLYN-assigned terminal numbers. Her observations were limited to public terminals in the university's graduate and undergraduate libraries. We scheduled observation periods until we collected between four and five hundred subject queries.

Although the format of transaction log data varied from system to system, the content was quite similar. Computer programs printed transaction logs containing seven columns of data to enable the researchers to interpret user-system activity. Figure 2.1 shows a portion of user-system activity from the SULIRS online catalog. An explanation of the seven columns of data follows.

Starting on the far left is a column of line numbers in numerical order that the computer program generates for all user-system activity involving the user's entry of a subject-access point. The second column gives the date by reserving two numerical positions for each of the following: (1) year, (2) month, and (3) day. The third column is a time stamp consisting of two numerical positions for each of the following: (1) hour, (2) minute, and (3) second. The fourth column reserves three alpha-numerical positions for identifying the terminal at which the activity took place. The fifth column gives one or more two-letter SULIRS commands entered by the user. An example of one command is shown on line number 46 in which the "su" command directed the system to search for the subject query "smoking woman" in title and/or subject heading fields. An example of two two-letter commands is shown on line number 57 in which the user's entry of "au" and "wd" commands directed the system to search and find the subject query "language" in author fields and in title and/or subject heading fields. The seventh column lists the subject queries users entered into SULIRS. The sixth column reports the number of bibliographic records SULIRS retrieved in response to the subject queries that follow in the seventh column.

```
(1) (2)      (3)      (4) (5)      (6)   (7)
41  880426   121525   C17 sb       00483 motivation
42  880426   121548   C17 sb       00005 organizational
                                         motivation
43  880426   124320   C17 sb       00000 ford motor co.
44  880426   124345   C17 sb       00248 ford
45  880426   124517   C17 sb       00006 ford automobile company
46  880426   125417   C17 su       00000 smoking woman
47  880426   125555   C17 su       00001 effects of smoking
48  880426   125638   C17 su       00004 tobacco women
```

```
49  880426   125733   C17  su      00000  smoking fetus
50  880426   125754   C17  su      00003  tobacco fetus
51  880426   132946   C17  sb      00000  auther james balfour
                                          speeches on zionism
52  880426   135523   C17  su      00000  reading integrated with
                                          writing
53  880426   135539   C17  su      01575  writing
54  880426   135634   C17  su      00085  writing/reading
55  880426   140405   C17  su      00096  antiques
56  880426   141320   C17  ausb    00002  linguistics
57  880426   141529   C17  auwd    00006  language
58  880426   141940   C17  sb      00005  motifs art deco
59  880426   142234   C17  sb      00003  motifs art nouveau
60  880426   142733   C17  sb      00013  medical encyclopedia
61  880426   143914   C17  sb      00000  delacroix and colr
62  880426   143929   C17  sb      0044   gaugin
...
```

Key:
(1) line number
(2) date of transaction
(3) time stamp
(4) terminal identifier
(5) user-entered command
(6) number of bibliographic records
(7) user-entered query

Figure 2.1. Portion of a SULIRS log

We limited our study to the *initial access points* in subject searches. The decision to study initial subject access points in subject searches was guided by the objective of the original empirical study, viz. to determine how online systems could respond to user queries with the subject searching approach most likely to succeed in providing relevant information (Drabenstott and Vizine-Goetz 1994, 141). If this objective was achieved, then subsequent queries users enter might not be necessary.

We scanned printed transaction logs of user activity and demarcated subject searches. Demarcation was based on the period of time between sequentially-numbered transactions. It was also based on the context and subject matter of a query with those preceding and following. Determining when one search ended and the next one began was sometimes difficult. When in doubt, we considered the user query a new search.

In the sample SULIRS log (figure 2.1), we would have demarcated ten searches starting on lines 41, 43, 46, 51, 52, 55, 56, 58, 60, and 61. Queries printed on lines 41 through 45 could be the same search, i.e., “Organizational motivation at the Ford Motor Company.” However, the time stamp indicated that the query on line 43, i.e., “ford motor co.” was entered into SULIRS almost thirty minutes after the preceding access point on line 42, i.e., “organizational motivation.” Thus, we determined that queries on lines 41 and 42 were entered in a search separate from queries on line 43 through 45. On occasion, we encountered searches on the same topic, sometimes using the same words, that were entered into the same terminal but separated by one or two hours. We considered these separate searches and analyzed them separately.

We sought answers to five general questions about the initial subject access points in searches:

1. Is this initial access point a legitimate subject term?
2. What subject searching capability would the existing search-tree configuration invoke?
3. Would retrievals be satisfactory?
4. In the case of unsatisfactory retrievals, what subject searching capability would produce more satisfactory retrievals?
5. In the case of no retrievals, how could the search trees manipulate user queries to produce satisfactory retrievals?

2.3Subject Searching in the Four Selected Online Catalogs

2.3.1 Introduction

For the empirical study described in this report, we extracted subject queries from the online catalog transaction logs of four university research libraries: (1) SULIRS at Syracuse University, (2) ORION at the University of California, Los Angeles (UCLA), (3) LS/2000 at the University of Kentucky, and (4) MIRLYN at the University of Michigan. Table 2.1 lists the dates, number of initial subject access points, and number of terminals per library.

Table 2.1. Transaction Log Data

System	Logging dates	No. of terminals	No. of access points
SULIRS	April 23–May 10, 1988	14	571
ORION	June 4–14, 1988	33	508
LS/2000	Jan. 31, 1989–Feb. 20, 1989	11	418
MIRLYN	Feb. 10–17, 1993, Mar. 28–Apr. 5, 1993	14	419

All queries were extracted from public terminals; queries were not extracted from terminals searched by library staff. Almost six hundred access points were selected from SULIRS logs. SULIRS logging dates covered a time period that was the concluding third of the university's winter semester and included examination week. Queries were collected from ORION in the second half of the university's spring quarter in the two weeks that immediately preceded the university's examination week. Subject access points from LS/2000 numbered over four hundred and were collected from terminals during a three-week period shortly after the university's winter semester began in 1989. Queries were collected from MIRLYN in the week prior to spring break in winter 1993 and in the week preceding the university's examination week in the same semester.

This empirical study of user queries focused on defining the optimum response of online catalogs to the variety of subject terms entered by online catalog users. For this reason, we sought transaction logs from online catalogs that offered different approaches to subject searching in the hopes of obtaining a wide variety of subject access points. The only results that this study covered regarding the success of subject searches conducted in SULIRS, ORION, LS/2000, and MIRLYN were the numbers of bibliographic records retrieved in SULIRS, the number of assigned subject headings retrieved in ORION, and whether MIRLYN searches retrieved or failed to retrieve titles or subject headings in response to user queries.

Descriptions of subject searching in SULIRS, ORION, and LS/2000 were given in the original empirical study (Drabenstott and Vizine-Goetz 1994, 134–8); descriptions included sample subject searches (Drabenstott and Vizine-Goetz 1994, 343–51). Brief descriptions of subject searching in the four catalogs follow.

2.3.2 SULIRS

SULIRS featured a command-based interface and offered keyword, implicit and explicit Boolean searching. SULIRS searchers entered subject access points using the commands “lc” or “sb.” When users entered “lc,” SULIRS searched for the user-entered queries in subject heading fields of bibliographic records. When they entered “sb,” SULIRS searched for the words in user-entered queries in title and subject heading fields of bibliographic records. (SULIRS also accepted “wd” or “su” as equivalent forms of “sb.”)

When users entered queries bearing two or more words, SULIRS performed an implicit Boolean “and” operation. Users could also explicitly enter Boolean operators. For example, in response to the queries “sb;surrealist artist” or “sb;surrealist and artist,” SULIRS looked for the word “surrealist” in the title and/or subject heading fields of bibliographic records, looked for the word “artist” in the title and/or subject heading fields of bibliographic records, found records common to both words, and displayed to users the number of records retrieved and the option to display them. Subject searching in SULIRS was a one-step approach in which the system's retrieval of bibliographic records led directly to the display of retrieved bibliographic records bearing the terms in user queries.

For computer and manual analyses, we extracted subject access points from SULIRS transaction logs that users entered with the commands “lc,” “sb,” “su,” or “wd.” We also extracted subject access points that were entered using one or more of these commands with title, author, and/or series commands (see line numbers 56 and 57 in figure 2.1). When SULIRS retrieved more bibliographic records than the display buffer could hold (usually about one thousand bibliographic records), SULIRS informed users that they retrieved too many records and must reduce the number to display retrieved records. When the logs were restructured by project staff, the number 99,999 was written to logs for these high-posted searches even though the actual number of retrieved bibliographic records might have been lower.

2.3.3 ORION

ORION also featured a command-based interface and offered keyword, implicit Boolean searching. The major difference between the keyword searching of SULIRS and ORION was that the latter system retrieved an intermediary display of single assigned subject headings bearing the words in user queries. ORION users scanned the list of retrieved headings, selected the one(s) that interested them, and then retrieved the bibliographic records bearing the selected heading. ORION users entered their access points using the command "browse" (or the shortened form "b") followed by the field label "su" and their terms. For example, in response to the user query "b su stress," ORION searched for this word in the main headings and subdivisions of assigned subject headings and retrieved assigned subject headings such as "Job stress" and "Metals — Stress corrosion."

When users entered more than one word following the "browse su" command, ORION performed an implicit Boolean "and" operation. For example, ORION interpreted the subject access point "b su bibliography on stress" as "bibliography and stress." It deleted "on" as a stopword, then looked for the word "bibliography" in assigned subject headings, looked for the word "stress" in assigned subject headings, found headings bearing both words, and informed users of the number of headings retrieved and gave them the option to display them. Subject searching in ORION was a two-step approach in which the system responded to the retrieval of assigned subject headings with an intermediary list of single headings bearing words in the user query. The system encouraged users to browse and select listed headings that expressed their topics of interest to retrieve a summary list of the bibliographic records to which the selected headings were assigned.

ORION systems staff extracted subject access points from ORION transaction logs that were entered using the "browse" command and "subject" field label, i.e., "b su." They also extracted the number of assigned subject headings retrieved by subject access points.

Occasionally, ORION users entered personal names using the "browse" command and "subject" field label. Such access points were considered subject access points and included in both computer and manual analyses. The number of headings retrieved by such access points was usually zero or close to zero because ORION users had to use the system's "name" index to retrieve subject headings for personal names as subjects (and authors).

2.3.4 LS/2000

LS/2000 was the only catalog of the four catalogs studied that featured a menu-based interface. Library staff could choose the types of searches to offer to patrons, fields of bibliographic records for search and display, indexes for online browsing, and many other parameters. LS/2000 offered both alphabetical searching and keyword, implicit Boolean searching. Alphabetical searches displayed subject headings beginning with the same words as user queries. Keyword searches allowed users to enter search for one word at a time in various fields of bibliographic records and combine the results of separate searches for one-word queries.

LS/2000 converted end-user selections from menus into various commands and wrote these commands to transaction logs. We extracted subject access points from LS/2000 transaction logs that corresponded to the logged commands SG (Subject — Geographic), SJ (Subject Topical Authority), SK (Subject — Medical), SU (Subject), and KW (Keyword). Command SU corresponded to initial menu option 3, SK to initial menu option 4, and KW to entry of a keyword. Commands SG and SJ corresponded to secondary menu options 7 and 8, respectively, in the “Other Searches” menu. (Drabenstott and Vizine-Goetz (1994, 136–8) gave explanations on the different subject searches and commands LS/2000 made available to users.)

2.3.5 MIRLYN

MIRLYN had a command-based interface that allowed users to perform alphabetical searches for subject headings and keyword-in-record subject searches. MIRLYN responded to the user’s selection of the “s=” command with alphabetical searches and displayed a truncated screen bearing unsubdivided and subdivided subject headings beginning with the same words as the words in user queries. Alphabetical searching in MIRLYN was a two-step approach in which the system responded to the retrieval of assigned subject headings with an intermediary list of single headings bearing words in the user query. The system encouraged users to browse and select listed headings that expressed their topics of interest to retrieve a summary list of the bibliographic records to which the selected headings were assigned.

MIRLYN responded to the user’s selection of the “k=” command with the results of a keyword-in-record search for words in user queries. At Michigan, MIRLYN’s keyword-in-record search searched *all* fields of bibliographic records for words in user queries. That is, this search was not limited to retrievals from selected fields. Keyword-in-record searching in MIRLYN was a one-step approach in which the system’s retrieval of bibliographic records led directly to the display of retrieved bibliographic records bearing the terms in user queries.

MIRLYN systems staff extracted subject access points from MIRLYN transaction logs that were entered using the “s=” or “k=” command. They distinguished between low-and high-posted alphabetical (“s=”) searches by writing “ind” and “gui” to the postings field of transaction log records, respectively. For keyword-in-record searches, they wrote the number of retrieved bibliographic records to logs. In the data analysis, we converted these alpha-numeric results to zeros (for no retrievals) or ones (for one or more retrievals).

2.4Subject Searches and Initial Access Points

Table 2.2 gives the percentages of initial access points in legitimate subject-term subcategories and in a non-legitimate subcategory. Queries placed in the latter were discarded from subsequent analyses.

Table 2.2. Types of Initial Access Points

Categories	Total (n=1,916)	SULIRS (n=571)	ORION (n=508)	LS/2000 (n=418)	MIRLYN (n=419)
Topical	60.3	60.8	70.3	50.7	57.3
Personal name	11.4	13.1	5.9	15.8	11.5
Geographic	4.8	2.3	4.7	8.4	4.8
Corporate name	2.9	3.5	2.4	2.6	2.9
Topical and geographic	8.3	6.7	11.2	1.0	14.1
Topical and personal name	1.5	2.1	0.6	1.2	2.2
Topical and corporate name	0.2	0.5	0.0	0.0	0.0
Topical, personal name, and corporate name	0.2	0.2	0.0	0.0	0.5
Non-legitimate	10.4	10.8	4.9	20.3	6.7
Total	100.0	100.0	100.0	100.0	100.0

At all four data collection sites, the majority of user queries for subjects were topical subjects. Users searching SULIRS, LS/2000, and MIRLYN entered large percentages of subject queries for personal names. The small percentage (5.9%) of subject queries for personal names that ORION users entered were actually entered incorrectly. ORION required users to use the system's "find name" or "browse name" commands to search personal-name queries (see section 2.2.3). Users searching ORION and MIRLYN entered large percentages of subject queries bearing both topical and geographic name elements. Multiple-element queries, that is, queries bearing topical element(s) and one or more other element types, represented between two and seventeen percent of user queries for subjects. Non-legitimate queries ranged from a low of 4.9% of subject queries (in ORION) to a high of 20.3% of subject queries (in LS/2000). Section 2.5 focuses on non-legitimate queries and explains reasons why there were so many such queries in LS/2000.

Figure 2.2 summarizes the total percentages of types of initial access points across all four libraries.

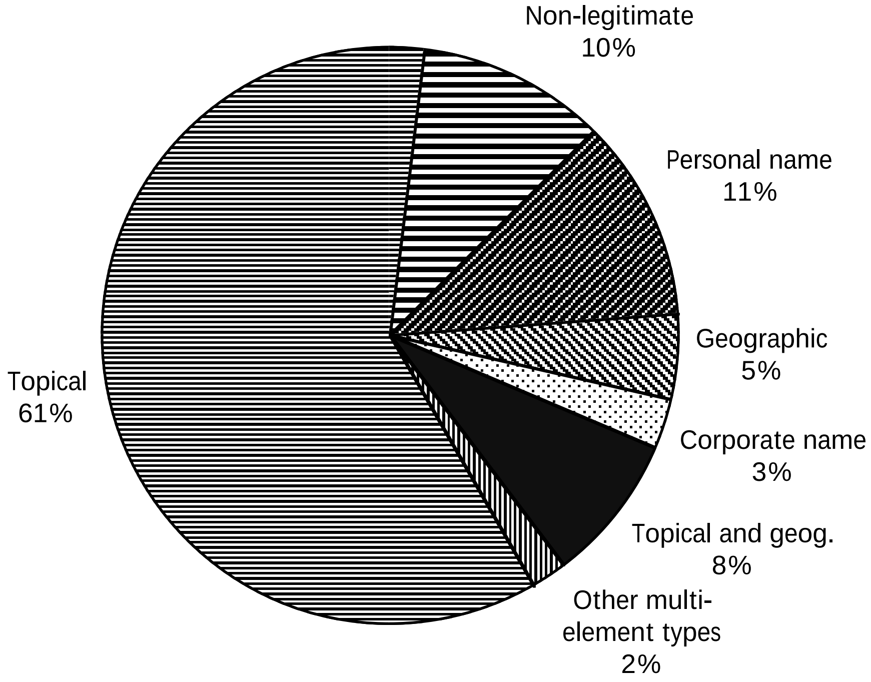


Figure 2.2. Types of initial access points

Overall, about three of every five queries contained only topical elements. Personal names accounted for 11% of user queries. The most frequent multi-element query contained topical and geographic elements and represented about 8% of user queries for subjects. Non-legitimate queries accounted for 10% of user queries for subjects.

Table 2.3 is a fact sheet with details about the total of 1,919 searches and initial access points in this study.

Table 2.3. Fact Sheet for Initial Access Points in Subject Searches

Facts	**Total**	**SULIRS**	**ORION**	**LS/2000**	**MIRLYN**
Searches					
No. of searches	1,919	571	511	418	419
No. of access points	5,823	1,524	1,637	1,318	1,344
Avg. no. of access points per search	3.0	2.7	3.2	3.2	3.2
Maximum no. of access pts. in a search	38	18	26	26	38
Initial access points					

Avg. no. of words per access point	1.9	2.0	1.8	1.5	2.2
Maximum no. of words in an access pt.	10	8	10	7	10
Avg. no. of retrievals per access point	N/A	205.5*	352.2†	N/A	N/A
Percentage of access points with zero retrievals	N/A	30.7	38.8	N/A	28.6
Percentage of access points with retrievals > 999	N/A	6.1*	5.8†	N/A	N/A
Avg. no. of words in access points for retrievals > 999	N/A	1.2*	1.1†	N/A	N/A
Avg. no. of access points in searches where initial access point has retrievals > 999	N/A	4.5*	4.9†	N/A	N/A

*Number of bibliographic records retrieved. This was an estimate based on substituting 3,000 retrievals for the 99,999 retrievals that was written to logs when number of retrievals exceeded about 1,000.
†Number of assigned subject headings retrieved.

Searches averaged three access points. The longest searches were two searches in ORION and LS/2000 containing twenty-six access points and one MIRLYN search containing 38 access points. The ORION user began with the access point "married people" and entered twenty-five more access points containing variants of the word "marriage" or words for the conditions that marriage brings, e.g., "love," "romance," "problems," and/or words about stages, e.g., "cycles," "lifecycles," "steps," "process." Phrases were also entered such as "marriage stages," "marriage processes," "marital steps," and "marital problems." The LS/2000 user began with the access point "career counseling" and entered twenty-five more access points describing related topics such as "high school guidance counseling," "college, choice of," and "student financial aid." The MIRLYN user began with the initial access point "macrobiotic diet" and entered several access points on eating disorders, e.g., "eating disorders," "eating disorders in children," "anorexia," "disordered eating," "disordered eating in women," and dieting, e.g., "fasting," "dieting," "reducing," "women and dieting," "women and diet," "diet pills," and "weight loss."

Initial access points averaged 1.9 words and ranged from a low of 1.5 words in LS/2000 to a high of 2.2 words in MIRLYN. Of the longest initial access points in the three systems, one has several spelling errors, one contains an addition/omission error, and one was probably entered by a user who was displaying a bibliographic record bearing a personal-name subject heading of interest:

- federal home loan bank system u s periodicalsiii9iu is iiiiiiiii (ORION, 10 words)
- women in the work force in the 1900–1950 (SULIRS, 8 words)
- legal cases involving life support systems in humans (SULIRS, 8 words)
- 2a concise history of the middle east (LS/2000, 7 words)
- vlad III prince of wallachia 1430 or 31–1447 or 7 (MIRLYN, 10 words)

Comparisons between systems with respect to number of retrievals were difficult to make because the three systems that reported number of retrievals counted different things. SULIRS reported the *number of retrieved bibliographic records*. The number given in the fact sheet (Table 2.3) was an estimate based on substituting 3,000 retrievals for the number 99,999 that SULIRS reported to users when the number of retrievals exceeded about one thousand. ORION reported the *number of retrieved assigned subject headings*. MIRLYN distinguished between *types of subject heading displays* in alphabetical searches and reported the *number of retrieved bibliographic records* in keyword-in-record searches.

Initial access points entered by SULIRS users *averaged over two hundred retrievals*. ORION users retrieved an *average of about 350 assigned subject headings* in response to their initial access points. About three in every ten initial access points produced zero retrievals in SULIRS (175 of 570, or 30.7%). Almost four of every ten initial access points produced zero retrievals in ORION (200 of 515, 38.8%). A little over one-quarter of MIRLYN searches failed to produce retrievals. The initial access points in thirty-five SULIRS searches and thirty ORION searches produced one thousand or more retrievals. In both SULIRS and ORION, access points that produced one thousand or more retrievals were usually composed of one word; this was far below the average of three words per access point. Users whose initial access points were so high posted entered between four and five access points per search in comparison to the average of three access points per search.

2.5Initial Access Points Discarded from the Analyses

2.5.1 Introduction

We discarded known-item and other access points from the data analyses. Prior to discarding them, we placed them in several categories for known-item searches and other access points. Figure 2.3 summarizes the percentages of discarded initial access points per category.

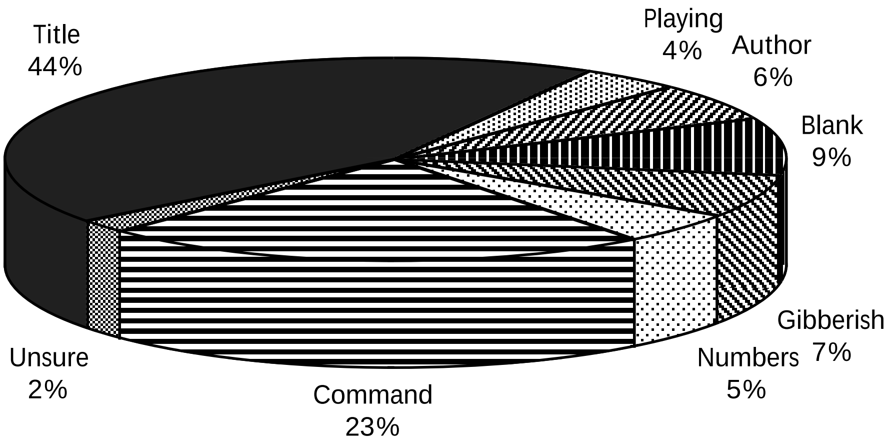


Figure 2.3. Discarded initial access points

Half (50%) of discarded initial access points were known-item searches. Most (44%) were title searches. The rest (6%) were author or author-title searches. Commands accounted for 23% of discarded initial access points. Included in this category were help commands, LS/2000 remote signon command, and other commands. Five percent of discarded access points were numbers; these number searches included dates and call numbers. Gibberish accounted for 7% of discarded access points and characterized access points with meaningless terms composed of two or more of more of the following: letters, numbers, or punctuation. Queries placed in the “Playing” category included sex words, swear words, and satanic words.

Table 2.4 is a fact sheet with details about the total of 203 searches and initial access points discarded from this study.

Table 2.4. Fact Sheet for Discarded Initial Access Points

Facts	**Total**	**SULIRS**	**ORION**	**LS/2000**	**MIRLYN**
Searches					
No. of searches	203	62	28	85	28
No. of access points	361	108	51	155	47
Avg. no. of access points per search	1.8	1.7	1.8	1.8	1.7
Maximum no. of access pts. in a search	15	10	6	15	6
Initial access points					
Avg. no. of words per access point	2.0	1.7	2.9	1.6	2.9
Maximum no. of words in an access pt.	8	6	8	8	8

Avg. no. of retrievals per access point	N/A	282.3*	261.8†	N/A	N/A
Percentage of access points with zero retrievals	N/A	58.1	39.3	N/A	39.2
Percentage of access points with retrievals > 999	N/A	8.1*	7.1†	N/A	N/A

*Number of bibliographic records retrieved. This was an estimate based on substituting 3,000 retrievals for the 99,999 retrievals that was written to logs when number of retrievals exceeded about 1,000.
†Number of assigned subject headings retrieved.

A little over two hundred searches were discarded. Discarded searches averaged 1.8 access points per search which was somewhat shorter than the average number of 3.0 access points per search for initial access points generally (see Table 2.3). Although these access points retrieved bibliographic records, percentages of access points that failed to yield retrievals were quite a bit larger for ORION and MIRLYN than percentages for initial access points generally (see Table 2.3).

2.5.2 Known-item Access Points

Queries formulated like a book or journal title and entered into the three systems studied using their respective subject search commands were categorized as known-item searches. Sometimes we had to search one or more online catalogs to verify whether the query actually retrieved a specific known-item. Examples of known-item searches are:

- bartlett's quotations (SULIRS)
- dubliners (LS/2000)
- fireside chats (LS/2000)
- lcsh (ORION)
- journal of basic enginerring [sic] (ORION)
- robinson rusoe [sic] (ORION)
- ti=social psychology of fashion (LS/2000)
- two essays on analytical psychology (SULIRS)
- a ten year national highway program (MIRLYN)
- wild palms and the old man (MIRLYN)

Nearly all known-item searches were title searches. Examples of author-title searches are "auther [sic] james balfour speeches on zionism" (SULIRS), "cervantes saavedra, miguel de. don quixote" (MIRLYN), and "cook zeus" (MIRLYN). Three access points entered into LS/2000 and MIRLYN included the system's author search command with the name, i.e., "au=adams, ansel" (LS/2000), "au=fujikake" (LS/2000), and "tour.au." (MIRLYN), and, thus, were categorized as known-item searches. Users must have selected a subject search option from the initial menu, then entered names with an explicit author command, i.e. “au=” or “.au.” In LS/2000, users must first choose a subject search option; consequently, the system treated the query including author command as a subject term.

The transaction logs we examined only listed queries users entered using *subject* commands available in the three systems. Users probably entered known-item access points using known-item search commands. If such searches failed to retrieve items, users might have reentered queries using a subject search command. In another matching study (Markey 1984, 65–73), the researcher examined unabridged transaction logs and noticed that users reentered the same query using several different subject search commands and known-item commands. For example, “surrealist painters” was reentered using commands for subject, title, and author searches. When users entered queries using improper search commands, they might have been looking for a manageable number of retrievals or using another strategy following one that produced zero retrievals. Additionally, they might not have understood how to use the particular system’s different searching commands in view of their information needs.

2.5.3 Other Discarded Initial Access Points

Six different types of initial access points were defined after access points were categorized in the “other” access points category. Types and examples of access points follow. Several of them were connected with the search formulation and operational features of the three online catalogs. Percentages of queries are also cited.

1. Unsure (2% of other access points).

This category includes word stems, e.g., “ger,” “muc,” “per,” “dans,” and incomplete phrases, e.g., “green p.” It also includes terms that the investigators did not understand, e.g., “vignaux,” “maldef,” “sol gel,” “gabins,” were unable to verify, e.g., “cort,” “cort thinking program,” “brazil bragantina,” “rietveld,” “group f/64,” “framework II,” “lsi logic corporation,” “parens patriae,” or did not give enough information to enable them to categorize the query, e.g., “directory,” “pilates,” “interpretations.” Queries categorized here were taken from all but the MIRLYN system.

2. Blank line (9% of other access points).

Blank lines were entered as access points into all but the MIRLYN system. Entry of blank lines could have been inadvertent or an attempt on the part of the user to get help or induce the system to respond with the ready prompt.

3. Commands (23% of other access points).

Occurrences of the SULIRS command name “s” sometimes composed entire queries. SULIRS asked users to begin a new search with the “s” command when the system was in the middle of a search, i.e., displaying retrieved citations or reporting the number of citations retrieved. Users’ entry of the “s” was probably inadvertent.

To override menus, LS/2000 users could enter search commands directly. Sometimes they entered commands when LS/2000 was expecting a query. Thus, LS/2000 performed a search for the command name entered. Other times they entered incorrect command names with correct command syntax. Examples are “antianri/es,” “ls/2000/es,” “-/AT,” “eses,” “3/ti.”

Remote signon commands entered as subject queries came exclusively from University of Kentucky transaction logs. “Liv” was the remote signon command to LS/2000 at Kentucky. When online catalog users found an LS/2000 terminal that had been turned off or was in the middle of the last user’s search, they sometimes entered the “liv” command in an attempt to restart the system. Instead, LS/2000 resumed operation from the point when the terminal was turned off or the last user left the terminal. If LS/2000 was in subject search mode, it performed a subject search for the user-entered query “liv.” Nearly half of LS/2000 queries placed in the “other access point” category were also placed in this subcategory.

One SULIRS user entered “help” as his initial query. The correct “help” command in SULIRS was the entry of a question mark (?).

4. Gibberish (7% of other access points).

Of the three systems, gibberish was most frequently encountered in LS/2000. Examples are “ba’ c,” “cu=508 r152 p p,” “p i,” and “jqp.”

Two SULIRS queries were composed entirely of letters that looked like abbreviations, i.e., dwi and naevc. No heading or *see from* tracing in LCSH or assigned subject heading expressed the former. We were unable to verify the latter.

5. Numbers (5% of other access points).

LS/2000 users occasionally entered numbers into the system. When numbers ranged from 1 to 5, they were probably users’ inadvertent reentry of a number corresponding to a particular searching option from the initial menu. Other examples of numbers are “623,” “34343434,” and “31113131.”

We encountered dates as initial access points in SULIRS and LS/2000. Examples are “2019,” “1974,” and “1983.” When access points entered into LS/2000 produced too many retrievals, the system prompted users to reduce the number by date of publication or other criteria. Dates entered at this point would not be the initial access point in searches unless there was considerable time between the entry of the previous access point and the date that caused us to consider the date as a new search, and thus, an initial access point in a search. Additionally, LS/2000 users could have begun searches with dates because the system allowed users to reduce retrievals by entering dates. One SULIRS user entered “la227.3.a43” using the system’s “lc” command. This subject search was restricted to finding words in LC subject headings. SULIRS searched call numbers preceded by the “cn” command.

6. Playing (4% of other access points).

An example of a query categorized as playing was "peter" (followed by "francine"). Often the context of such queries helped the investigators categorize them here.

We also used context to determine whether sex words should be included or discarded from the study. When sex words started and comprised an entire search, we categorized them here and discarded them from subsequent analyses. Sex words that were exact or partial matches of LCSH-mr or assigned subject headings were included in the analysis. Examples of such queries are "sex," "penis expansion," "erotica," "anal sex." Reviewing lists of access points, we noted that sex words were more likely to occur in the middle or at the end of a search. Perhaps users entered such words as a diversion during a long search or as an expression of frustration when their searches were not particularly fruitful.

Context also played an important role in identifying satanic words. When such words started and comprised an entire search, they were categorized here, and, thus, discarded from subsequent analyses. The one example is "666." Satanic words that were exact or partial matches of headings or *see from* tracings in LCSH or assigned subject headings were included in the analysis. The one example of the latter is "satan."

2.6Chapter Summary

Chapter 2 is an orientation to the subject queries users entered into online catalogs. It begins by describing the manual methods we used to analyze subject queries from four libraries' transaction logs and includes a description of transaction log formats (section 2.2). Accompanying this description are brief statements of subject searching in the four libraries' online catalogs (section 2.3).

For the empirical study described in this report, we extracted subject searches from the online catalog transaction logs of four university research libraries. We extracted a total of between four hundred and six hundred searches per participating library (Table 2.1) and sought answers to five general questions about the initial access points in subject searches:

1. Is this initial access point a legitimate subject term?
2. What subject searching capability would the existing search-tree configuration invoke?
3. Would retrievals be satisfactory?
4. In the case of unsatisfactory retrievals, what subject searching capability would produce more satisfactory retrievals?
5. In the case of no retrievals, how could the search trees manipulate user queries to produce satisfactory retrievals?

We extracted a total of 1,919 initial access points from the four libraries' transaction logs and placed them in legitimate subject-term categories and non-legitimate categories (section 2.4).

The former were queries composed of elements for topical subjects, geographic names, personal names, corporate names, and combinations of subjects and names (Table 2.2). The latter numbered 203 access points (Table 2.4). Most were known-item searches; other types were command names, gibberish, blank lines, and numbers (figure 2.3). Non-legitimate queries were discarded from subsequent analyses.

2.7 References

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

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

3 Search Trees for Subject Searching

3.1 Introduction

The search trees described in this chapter were the result of an empirical study of end-user queries for subjects. Details of the empirical study were summarized by Drabenstott and Vizine-Goetz (1994, 151–240).

Although search trees were meant for implementation in operational online catalogs, we used the search-tree configurations described in this chapter to determine the extent to which user queries produced and failed to produce retrievals using this configuration. Queries that failed to make matches were selected for further study. These queries were especially challenging because they resulted in enhancements to the existing search-tree configuration.

3.2 Search Tree Design

The designers of the Okapi online catalog first introduced search trees. They defined search trees as "a set of paths with branches or choices, which enable systems to carry out the most sensible search approach at each stage of the search" (Mitev, Venner, and Walker 1985, 94).

Six search trees for subject searching are presented in this chapter. Search trees implemented in Okapi were different from the search trees presented in this chapter for two important reasons. First, only half of the bibliographic records in Okapi's database contained subject headings. When formulating search trees, the Okapi designers did not place as much emphasis on subject headings as on other subject information that was present in most of the system's bibliographic records (Mitev, Venner, and Walker 1985, 40). The vast majority of bibliographic records created by American libraries are assigned subject headings based on the Library of Congress Subject Headings system (O'Neill and Aluri 1979). The search trees presented in this chapter were recommended for subject searching of bibliographic records assigned headings based on this system.

The second reason why our search trees differ from those of the Okapi designers stems from how the trees were developed. Okapi's designers noted that the sequence of search functions or approaches “evolved through a process of discussion and trial and error” (Mitev, Venner, and Walker 1985, 94). The search trees presented in this chapter were derived from the findings of an empirical study of the subject terms users enter into online catalogs (Drabenstott and Vizine-Goetz 1994, 151–240).

3.3 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 3.1 is the initial search tree.

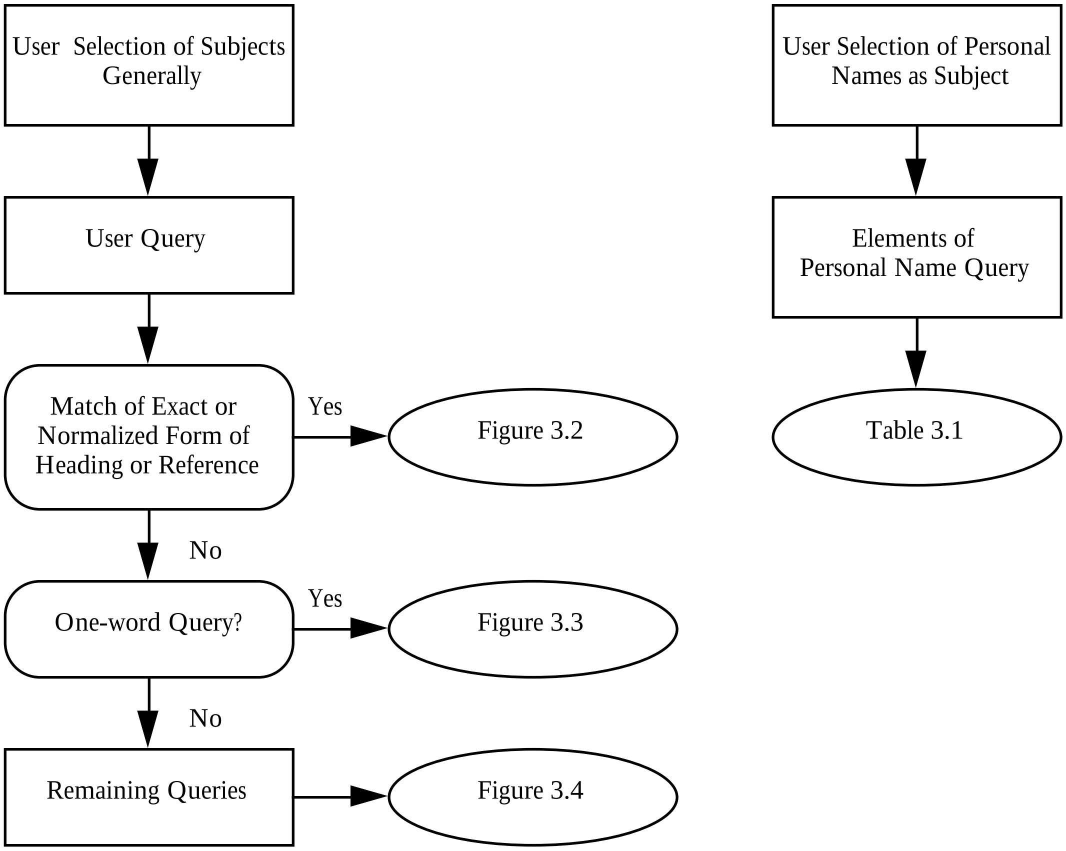

User Selection of Subjects Generally
User Query
Match of Exact or Normalized Form of Heading or Reference
Yes
Figure 3.2
No
One-word Query?
Yes
Figure 3.3
No
Remaining Queries
Figure 3.4
User Selection of Personal Names as Subject
Elements of Personal Name Query
Table 3.1

Figure 3.1. Initial search tree

In an operational system, the initial search tree would let users distinguish their queries for personal subjects from queries for topical subjects generally and, based on summary characteristics that systems determine about the latter types of user queries, dispatch them to a particular search tree that favors certain subject searching approaches over others.

In the analysis of end-user queries, all user queries for subjects generally were candidates for the exact approach. To effect an exact match, we manipulated user queries in the same way as controlled vocabulary terms would be 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 3.2). We recorded exact matches on tally sheets and also distinguished between various types of exact matches, e.g., exact matches, normalized matches, matches with spelling errors.

Unable to find an exact match, we chose 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 3.3). 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 3.4 and 3.5, respectively).

A system governed by search trees would prompt users entering personal name queries for the name elements and topic elements of personal name queries. Personal name queries would be handled by a search tree that was separate from search trees for subjects generally (Table 3.1). Systems would choose between alphabetical and keyword approaches depending on the type(s) of elements users enter and the ability of these elements to produce retrievals.

3.4 Search Trees for Subject Queries Generally

When user queries for subjects generally matched exact or normalized forms of controlled vocabulary terms, we recorded the exact approach to tally sheets as the subject search that made matches. Figures 3.2A and 3.2B depict the search tree that features the exact approach.

The exact approach search tree was split into two parts. Figure 3.2A depicts major events of the exact approach in which the system presents a conceptual map to subdivided forms of the matched heading, and, if available, gives users the option to browse related terms and other information about the matched heading.

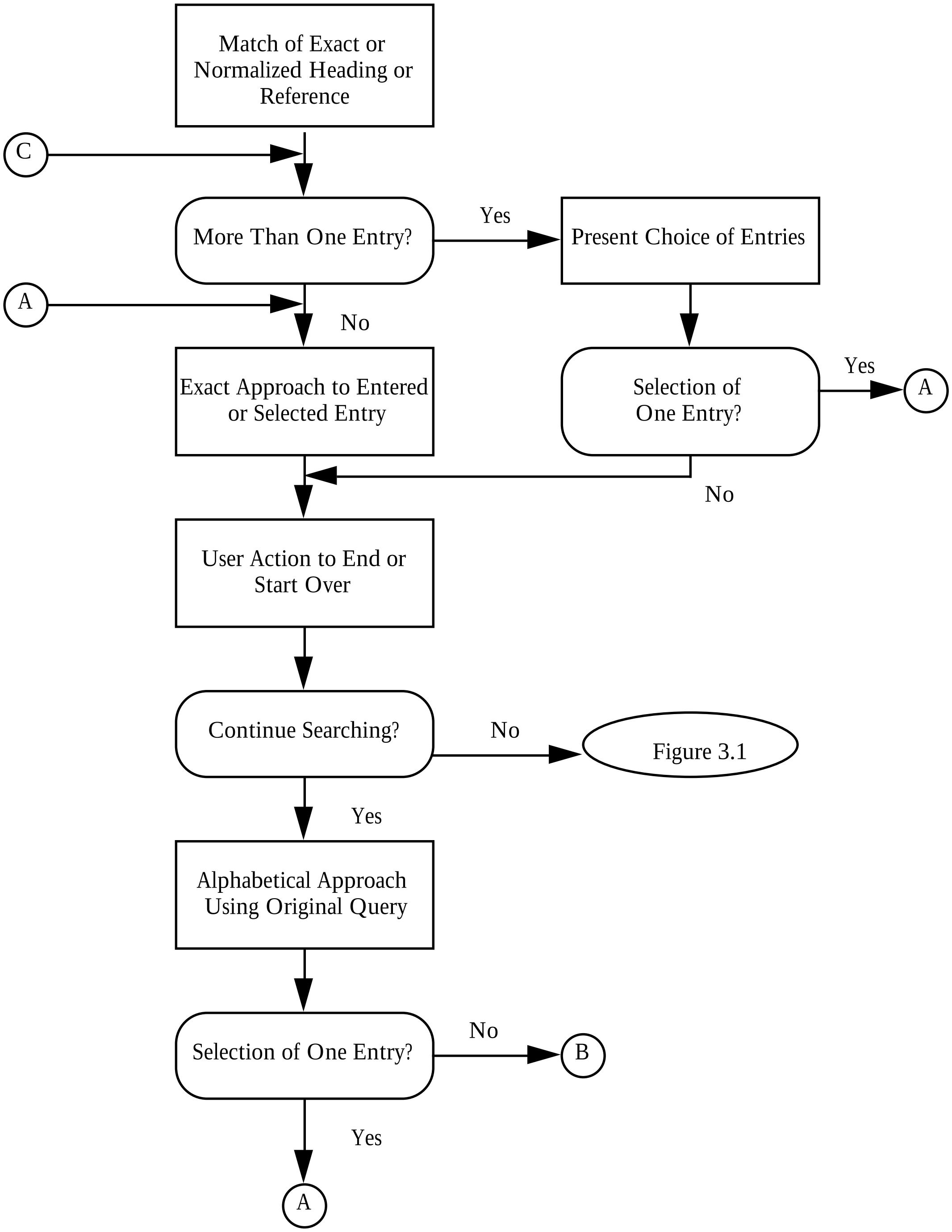

Match of Exact or Normalized Heading or Reference
C
More Than One Entry?
Yes
Present Choice of Entries
A
No
Exact Approach to Entered or Selected Entry
Selection of One Entry?
Yes
A
No
User Action to End or Start Over
Continue Searching?
No
Figure 3.1
Yes
Alphabetical Approach Using Original Query
Selection of One Entry?
No
B
Yes
A

Figure 3.2A. Search tree for the exact approach

Figures 3.2A and 3.2B depict major events following a user action to start over or end a search that begins with the exact approach. If users make an action to end the search or start over, systems would prompt them to continue searching. Users who continue could benefit from the search results of other approaches beginning with controlled vocabulary approaches. The search tree would submit their original query to the alphabetical approach and various keyword approaches beginning with the keyword-in-main-heading search. Thus, users whose queries invoked the exact approach have several other opportunities to retrieve additional titles.

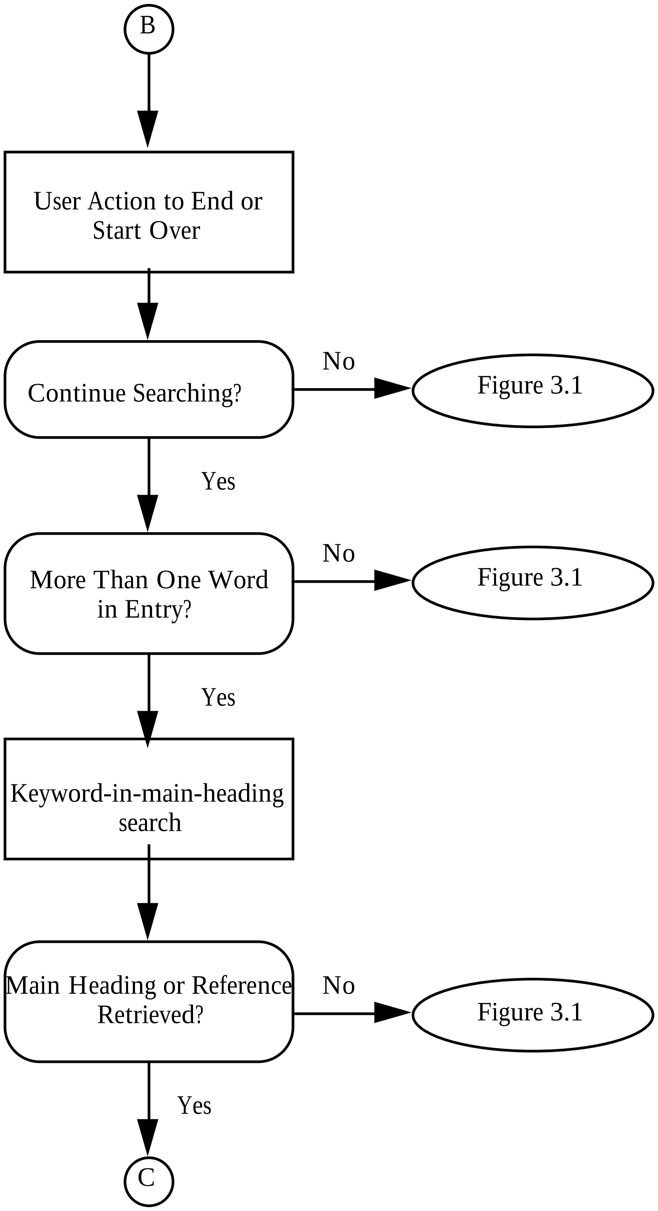

B
User Action to End or Start Over
Continue Searching?
No
Figure 3.1
Yes
More Than One Word in Entry?
No
Figure 3.1
Yes
Keyword-in-main-heading search
Main Heading or Reference Retrieved?
No
Figure 3.1
Yes
C

Figure 3.2B. Search tree for the exact approach (contd.)

Figure 3.2B 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.

The search tree for one-word queries is given in figure 3.3. 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, we checked their spelling and corrected them if necessary. We then submitted corrected queries to the initial search tree (figure 3.1) because revised queries might have matched exact or normalized forms of controlled vocabulary terms following the spelling correction.

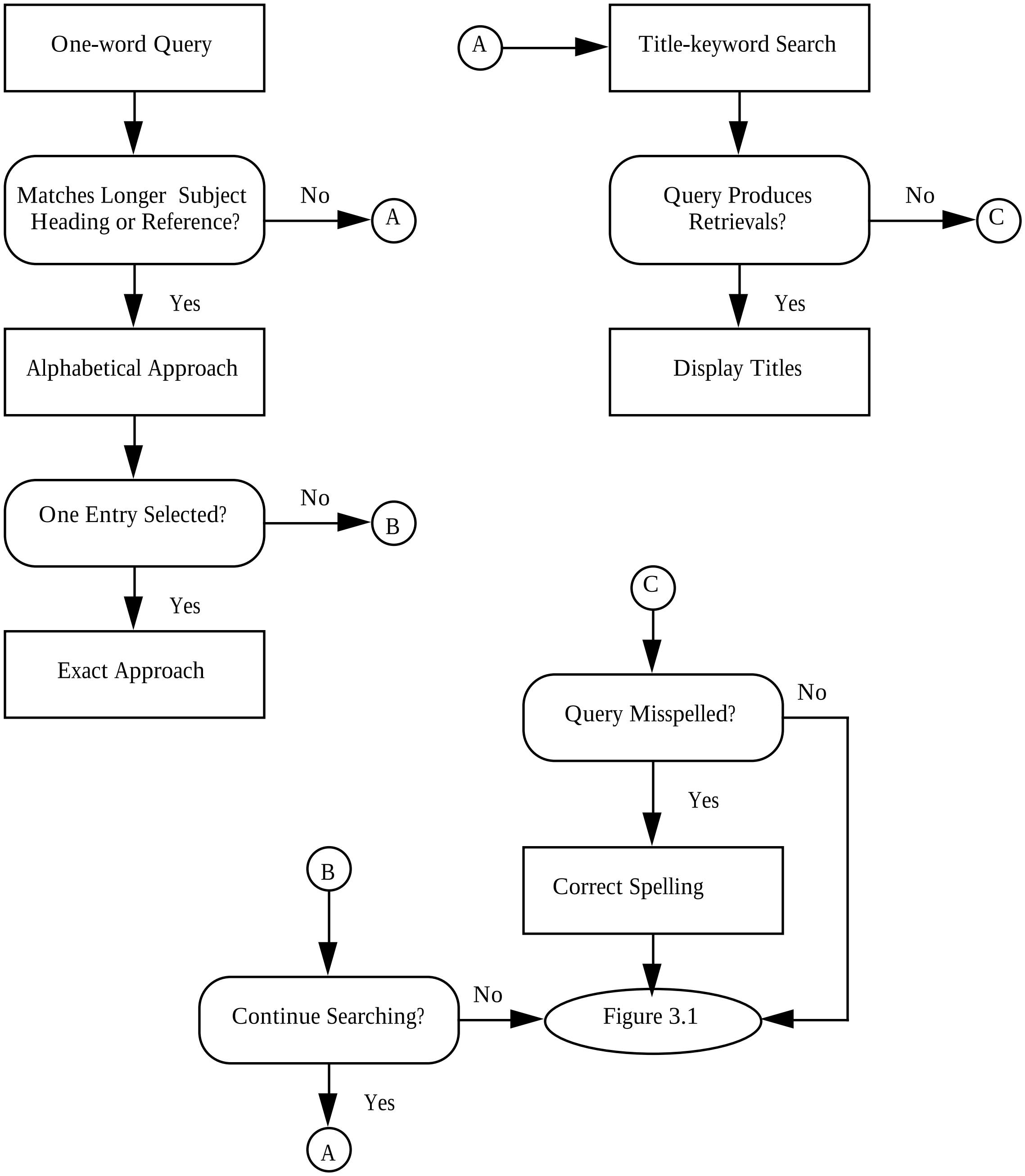


Figure 3.3. Search tree for one-word queries

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 would be candidates for the alphabetical approach (figure 3.4). The rest were submitted to a series of keyword searches that began with the keyword-in-main-heading search (figures 3.5A–3.5B). When the particular keyword search produced zero or too few retrievals, systems could continue searching using the next keyword approach in the series, thus, there would be ample opportunity for users to retrieve additional titles on their topics of interest.

Figure 3.4 is a search tree for queries composed of more than one word. Systems governed by search trees would respond with the alphabetical approach to queries that matched longer controlled vocabulary terms. To find additional material, systems would continue 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 heading “Civil rights movements.” The initial system response to such queries would be the alphabetical approach. When users responded to system prompts to continue searching using the original query, systems would continue with the results of title-keyword searches.

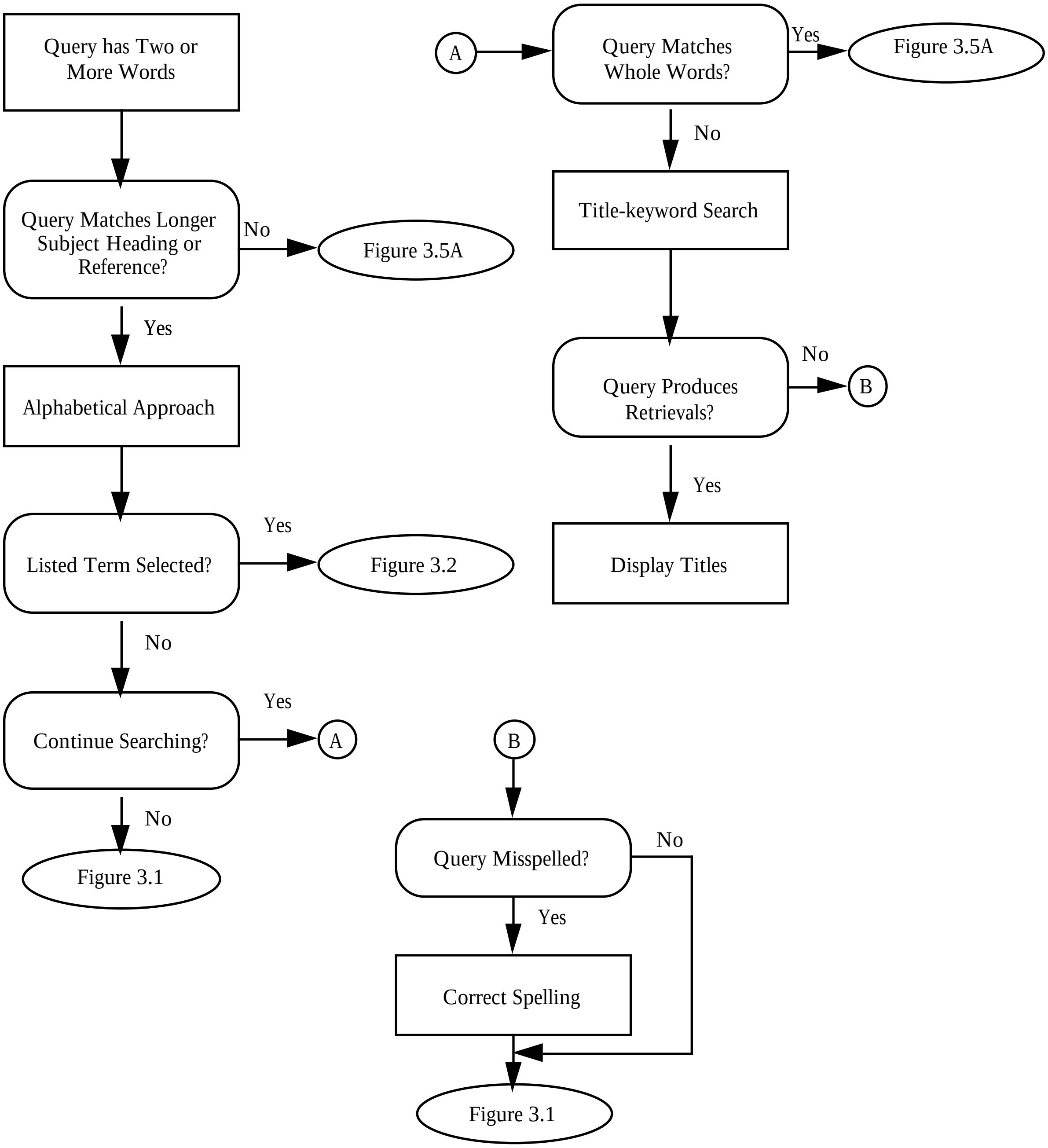


Figure 3.4. Search tree for multi-word queries featuring the alphabetical approach

We submitted the remaining queries composed of two or more words to a search tree for keyword approaches (figures 3.5A–3.5B). The search tree shown in figure 3.5A features the submission of queries to controlled vocabulary searches.

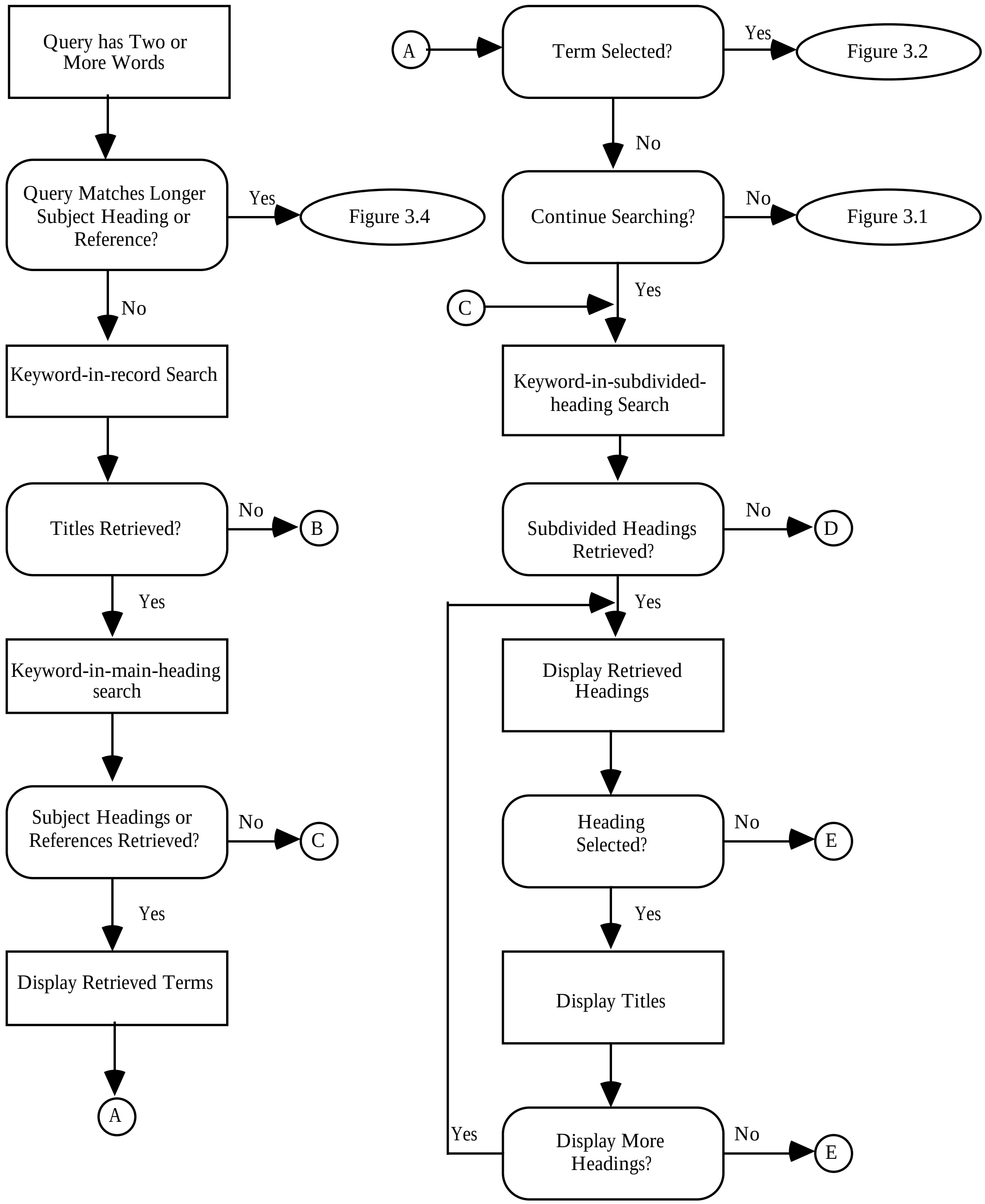

Query has Two or More Words
Query Matches Longer Subject Heading or Reference?
Yes
Figure 3.4
No
Keyword-in-record Search
Titles Retrieved?
No
B
Yes
Keyword-in-main-heading search
Subject Headings or References Retrieved?
No
C
Yes
Display Retrieved Terms
A
A
Term Selected?
Yes
Figure 3.2
No
Continue Searching?
No
Figure 3.1
Yes
C
Keyword-in-subdivided-heading Search
Subdivided Headings Retrieved?
No
D
Yes
Display Retrieved Headings
Heading Selected?
No
E
Yes
Display Titles
Display More Headings?
Yes
No
E

Figure 3.5A. Search tree for multi-word queries featuring keyword approaches

In an operational online catalog, systems would first use the keyword-in-record search to ensure that individual query words were posted in the catalog. Furthermore, they would perform a keyword-in-record search prior to submitting queries to controlled vocabulary searches. If one or more query words failed to produce retrievals, the entire series of keyword searches would fail. If queries failed, the system would ask users to check the spelling of their queries. If users made changes, systems would start 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, systems could try to correct spelling using automatic spelling correction algorithms. Otherwise, systems would have to ask users for a different query because their original one failed to produce retrievals using any keyword approach. If retrievals were produced through the initial keyword-in-record search, they would not be shown to users. Instead, systems would continue searching beginning with the keyword-in-main-heading search. The initial keyword-in-record search was meant to save systems additional steps searching for a query that did not retrieve any bibliographic records through the entire series of keyword searches.

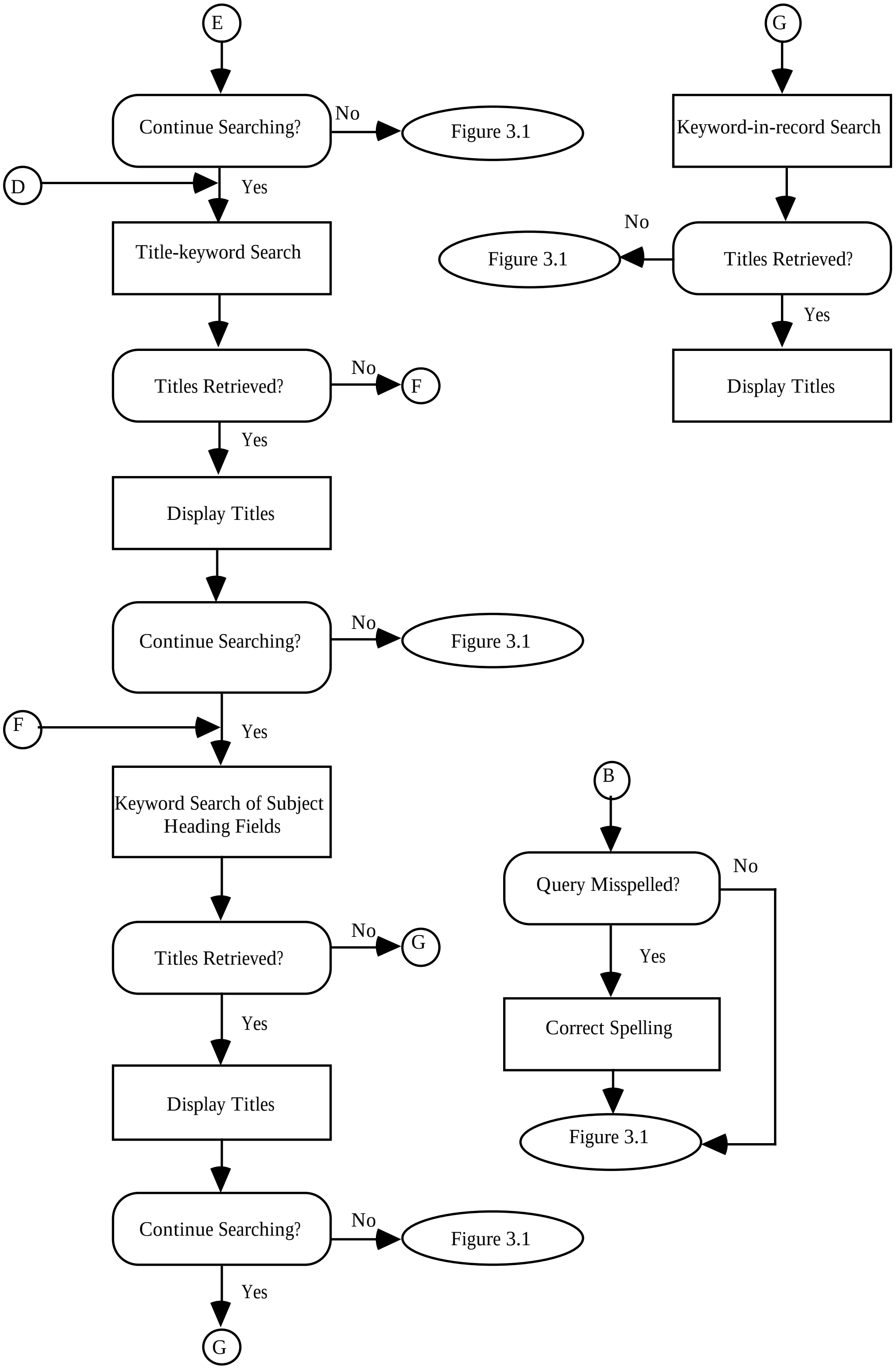


Figure 3.5B. Search tree for multi-word queries featuring keyword approaches (contd.)

Figure 3.5B 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. This search tree was the search tree "of last resort" for multi-word queries. Queries that failed to make matches through the searches on this tree were selected for further study. These queries were especially challenging because they resulted in enhancements to the existing search-tree configuration.

3.5 Search Tree for Personal-name Queries

Queries bearing elements for personal names were submitted to the search tree for personal-name queries. Since the four online catalogs from which we obtained transaction log data did not prompt users for the elements in personal-name queries, we identified the elements of personal-name queries on our own.

Table 3.1 lists the sequence of subject searching approaches to which elements of personal-name queries would be submitted in an operational system. If an approach failed to produce retrievals, we proceeded with the next approach and element(s) on the list to make matches.

Table 3.1. 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 us to single out queries containing topics. These queries would be 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 systems failed to find both elements, they 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.

3.6Chapter Summary

The search-tree configuration that was the result of an earlier empirical study of end-user queries figured prominently into the analysis of end-user queries in this study. We followed the sequence of subject searching approaches on the search trees to determine which approach would produce retrievals for end-user queries and which queries would fail to produce retrievals. Sections 4–6 describe the results of this analysis.

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 3.1). The search tree for matches of exact and normalized forms of controlled vocabulary terms favored the exact and alphabetical approaches (figure 3.2). One-word queries were given to a search tree that favored alphabetical or title-keyword searches (figure 3.3). Queries composed of two or more words matching longer controlled vocabulary terms were submitted to alphabetical and keyword-in-heading searches (figure 3.4). 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 3.5A and 3.5B).

User queries for personal names were controlled by a single search tree (Table 3.1). 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 they entered into online catalogs.

3.7 References

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

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

4 Search-tree Selections for End-user Queries

4.1Introduction

This chapter is a discussion of the characteristics of four types of matches of online catalog subject vocabulary: (1) exact, (2) partial, (3) keyword-in-heading, and (4) keyword matches. Each match type is further subdivided according to the particular elements in queries for subjects generally: (a) topical, (b) geographic names, (c) corporate names, and (d) combinations of topical subjects and these two types of names. Search trees can handle the four match types discussed in this chapter by submitting particular types to exact, alphabetical, keyword-in-heading, and keyword-in-record searches. Chapter 4 sets the stage for the discussion of non-matching queries and enhancements to the search trees to handle non-matching queries which follows in chapter 5.

4.2Facts on Queries for Subjects Generally

This section features fact sheets on the five types of queries for subjects generally: (1) topical subjects, (2) geographic names, (3) combinations of topical subjects and geographic names, (4) corporate names, and (5) combinations of topical subjects and corporate names.

4.2.1 Topical Subjects

Table 4.1 is a fact sheet with details about the total of 1,156 queries for topical subjects.

Table 4.1. Fact Sheet for Queries for Topical Subjects

Facts	Total	SULIRS	ORION	LS/2000	MIRLYN
Searches					
No. of searches	1,156	347	357	212	240
No. of access points	3,849	1,037	1,226	756	830
Avg. no. of access points per search	3.3	3.0	3.4	3.6	3.5
Maximum no. of access pts. in a search	38	18	26	26	38

Initial access points					
Avg. no. of words per access point	1.8	2.0	1.8	1.5	2.0
Maximum no. of words in an access pt.	8	8	5	6	6
Avg. no. of retrievals per access point	N/A	578.0*	329.6†	N/A	N/A
Percentage of access points with zero retrievals	N/A	26.5	56.0	N/A	24.2
Percentage of access points with retrievals > 999	N/A	6.9*	7.8†	N/A	N/A

*Number of bibliographic records retrieved. This was an estimate based on substituting 3,000 retrievals for the 99,999 retrievals that was written to logs when number of retrievals exceeded about 1,000.
†Number of assigned subject headings retrieved.

Queries for topical subjects averaged 3.3 access points per search which was a little longer than the 3.0 access points per search for initial access points generally (see Table 2.3). A few searches in three of the four systems were extremely long and contained over 25 access points. Over half of queries for topical subjects failed to produce retrievals through ORION's keyword-in-subdivided-heading search. About one-quarter of queries for topical subjects failed to produce retrievals in SULIRS and MIRLYN. When queries were posted, they retrieved many titles (578.0 titles in SULIRS) and subject headings (329.6 subject headings in ORION).

4.2.2 Geographic Names

Table 4.2 is a fact sheet with details about the total of 92 subject queries for geographic names.

Table 4.2. Fact Sheet for Subject Queries for Geographic Names

Facts	**Total**	**SULIRS**	**ORION**	**LS/2000**	**MIRLYN**
Searches					
No. of searches	92	13	24	35	20
No. of access points	292	28	58	141	65
Avg. no. of access points per search	3.2	2.2	2.4	4.0	3.3
Maximum no. of access pts. in a search	15	5	7	14	15
Initial access points					
Avg. no. of words per access point	1.3	1.3	1.4	1.2	1.3
Maximum no. of words in an access pt.	3	3	3	2	3
Avg. no. of retrievals per access point	N/A	1,106*	1,792†	N/A	N/A
Percentage of access points with zero retrievals	N/A	30.8	20.8	N/A	5.0
Percentage of access points with retrievals > 999	N/A	15.4*	29.2†	N/A	N/A

*Number of bibliographic records retrieved. This was an estimate based on substituting 3,000 retrievals for the 99,999 retrievals that was written to logs when number of retrievals exceeded about 1,000.
†Number of assigned subject headings retrieved.

There were many fewer subject queries for geographic names than for topical subjects. Queries for geographic names averaged 3.2 access points per search which was almost the same as the average for topical subjects (3.3). Searches were rather short and did not exceed 15 access points. Queries for geographic names were quite short. They averaged 1.3 words per access point. Most named countries. Examples of queries for geographic names are:

- hawaii
- iraq
- hong kong china
- great britain
- pakistan
- costa rica

Subject queries for geographic names were very highly posted. They retrieved many titles (1,106 titles in SULIRS) and subject headings (1,792 subject headings in ORION).

4.2.3 Combinations of Topical Subjects and Geographic Names

Table 4.3 is a fact sheet with details about the total of 158 subject queries for combinations of topical subjects and geographic names.

Table 4.3. Fact Sheet for Queries for Topical Subjects and Geographic Names

Facts	Total	SULIRS	ORION	LS/2000	MIRLYN
Searches					
No. of searches	158	38	57	4	59
No. of access points	551	88	227	16	220
Avg. no. of access points per search	3.5	2.3	4.0	4.0	3.7
Maximum no. of access pts. in a search	19	8	19	7	18
Initial access points					
Avg. no. of words per access point	2.9	2.8	2.8	1.8	3.1
Maximum no. of words in an access pt.	10	7	10	2	8
Avg. no. of retrievals per access point	N/A	287.6*	113.6†	N/A	N/A
Percentage of access points with zero retrievals	N/A	34.2	29.8	N/A	45.8

Percentage of access points with retrievals > 999	N/A	2.6*	1.8†	N/A	N/A

*Number of bibliographic records retrieved. This was an estimate based on substituting 3,000 retrievals for the 99,999 retrievals that was written to logs when number of retrievals exceeded about 1,000.
†Number of assigned subject headings retrieved.

Except at Syracuse, searches that began with an access point bearing both geographic and topical elements averaged about four access points. Access points bearing both elements averaged almost one word more than access points for these elements alone (Tables 4.1 and 4.2). Most access points contained at least two words, that is, a geographic name and topical subject consisting of one word each. Examples are "cinema sweden," "greek fiction," "french language," and "prague spring." Some were much longer like "tonkin gulf incidents, 1964," "regional planning and appalachian region," "labor unions and depression and michigan," and "atomic warfare and japan and history." Queries bearing both geographic and topical elements averaged far fewer titles (287.6) and subject headings (113.6) than queries consisting only of geographic-name elements (Table 4.2, 1,106 titles and 1,792 subject headings).

4.2.4 Corporate Names

Table 4.4 is a fact sheet with details about the total of 59 subject queries bearing corporate name elements. Four of the 59 subject queries contained corporate-name and topical-subject elements.

Table 4.4. Fact Sheet for Subject Queries for Corporate Names

Facts	Total	SULIRS	ORION	LS/2000	MIRLYN
Searches					
No. of searches	59	23	12	11	13
No. of access points	160	67	23	36	34
Avg. no. of access points per search	2.7	2.9	1.9	3.3	2.6
Maximum no. of access pts. in a search	12	12	5	7	6
Initial access points					
Avg. no. of words per access point	2.1	2.0	2.2	1.9	2.3
Maximum no. of words in an access pt.	6	4	5	6	5
Avg. no. of retrievals per access point	N/A	207.0*	0.1†	N/A	N/A
Percentage of access points with zero retrievals	N/A	43.5	91.7	N/A	38.5
Percentage of access points with retrievals > 999	N/A	4.4*	0.0	N/A	N/A

*Number of bibliographic records retrieved. This was an estimate based on substituting 3,000 retrievals for the 99,999 retrievals that was written to logs when number of retrievals exceeded about 1,000.
†Number of assigned subject headings retrieved.

Corporate-name access points initiated searches that averaged almost three access points per search. Access points contained about two words. Examples of corporate-name access points are “kmart,” “uaw,” “columbia university,” and “institute for sex research.” Few subject queries for corporate names resulted in retrievals in ORION. ORION users had to use the system’s “browse” command to search for corporate names. The users who entered the corporate names that we used in this study were using the system’s “find” command incorrectly.

4.3Queries for the Exact Approach

4.3.1 Introduction

Search trees for subjects generally first tested queries to determine if they were candidates for the exact approach (figure 3.1). Thus, our analysis of user queries for subjects generally began with this test. Table 4.5 is a fact sheet for user queries for subjects generally that systems governed by search trees would submit to the exact approach.

Table 4.5. Fact Sheet for Queries for the Exact Approach

Facts	Total	Topical	Geog.	Corp.	Topical-geog.
Searches					
No. of searches	832	653	79	40	60
No. of access points	2,660	2,081	255	108	216
Avg. no. of access points per search	3.2	3.2	3.2	2.7	3.6
Maximum no. of access pts. in a search	38	38	15	12	18
Initial access points					
Avg. no. of words per access point	1.6	1.5	1.2	2.1	2.8
Maximum no. of words in an access pt.	6	6	3	6	6
Percentage of access points with zero retrievals	13.8*	11.2*	8.3*	46.9*	23.3*
Percentage of access points with retrievals > 999	10.7†	15.3†	30.0†	0.0†	4.2†

*Does not include LS/2000
†Does not include LS/2000 and MIRLYN.

Exact matches accounted for 832 queries. This number was 43.4% of the total number of queries in the study. The majority (78.5%) of exact matches were queries for topical subjects. Exact matches figured into searches that averaged about three access points per search. One exact match initiated a search that featured 38 access points; this search began with the access point “macrobiotic diet” and included many other queries on dieting, food disorders, and substance abuse, e.g., “eating disorders,” “fasting,” “diets,” “alcoholic beverages,” “diet pills," and "amphetamines." Exact matches averaged between one and two words per query. A few queries were six words long. Examples are:

- united states food and drug administration
- titles of nobility and honor spain
- essential fatty acids in human nutrition

Queries for topical subjects and geographic names usually retrieved bibliographic records. Almost half of user queries for corporate names failed to retrieve titles even though they were exact matches of controlled vocabulary terms. High-posted searches were characteristic of queries that exactly matched topical subject headings or subject headings for geographic names.

4.3.2 Types of Exact Matches

Search trees for subjects generally first tested queries to determine if they were candidates for the exact approach (figure 3.1). Thus, our analysis of user queries for subjects generally began with this test. Table 4.6 shows the result. Of the total of 1,465 queries for subjects generally, 832 queries met the criteria for the exact approach.

Table 4.6. Types of Exact Matches

Type of exact match	Total (N=832)	Topical (N=653)	Geographic (N=79)	Corporate (N=40)	Topical-geographic (N=60)
Exact	62.9	67.7	65.8	17.5	36.7
Exact, spelling error	4.3	4.1	8.9	5.0	0.0
Exact, reference	13.0	13.3	2.5	37.5	6.7
Exact, spelling error, reference	0.8	0.5	0.0	10.0	0.0
Normalized	16.0	12.1	20.3	25.0	46.6
Normalized, spelling error	0.7	0.5	0.0	2.5	3.3
Normalized, reference	2.3	1.8	2.5	2.5	6.7
Total	100.0	100.0	100.0	100.0	100.0

In the exact matching process, we disregarded punctuation and capitalization. A little less than two-thirds of exact matches were exact matches of assigned subject headings overall. Another 18.1% of exact matches had spelling errors and/or matched cross references. References figured into 13.8% of exact matches overall. They figured into 47.5% of the various exact matches for corporate names.

In the normalization matching process, we disregarded punctuation, capitalization, word order, and stopwords. Normalized subject headings were exact matches of 16.0% of end-user queries overall. Normalized matches were particularly prevalent for topical-geographic query combinations where they accounted for almost two-thirds of matches.

A small fraction of exact matches of assigned subject headings are listed below:

israel	greek fiction	institute for sex research
nutrition	columbia university	latin language
astrology — history	hawaii	green movement
gourami	genocide	hate crimes
brothers and sisters	bosnia	drawings from photographs

On occasion, users entered queries that were placed into the exact-match category for assigned subject headings with spelling errors. Examples of such queries are:

User query	Matching subject heading
catholic churchu	Catholic church
viet nam	Vietnam
psibocylin	Psilocybin
guadalupe	Guadaloupe
austrailia	Australia
austrialalia	Australia
syracuse univeristy	Syracuse University
phptpgraphy	Photography

Misspelled exact matches of cross references were not common. One example is “3therapy” that was a cross reference under “Therapeutics.”

Less than twenty percent of exact matches were various matches of normalized terms. Table 4.7 lists user queries and various types of normalized matches.

Table 4.7. Normalized Matches

User query	Type of match	Matching term
korea south	normalized	Korea (South)
spoken in english	normalized, reference	Spoken English
evolution and frogs	normalized	Frogs — Evolution
maya	normalized	Maya (Hinduism)
arts fund raising	normalized, reference	Arts — Fund raising
mt.st.helens	normalized, reference	Mt. St. Helens (Wash.)
biocompatibilty+	normalized, spelling	Biocompatibility
taxonomy	normalized, reference	Taxonomy (Biology), Taxonomy (Botany), Taxonomy (Zoology)

4.3.3 Match Satisfaction

Exact matches did not guarantee that the matched headings and/or references would guide users to useful information on their topics of interest. Table 4.8 categorizes types of end-user queries by satisfaction categories.

Table 4.8. Satisfaction Categories for Exact Matches

Type of exact match	Total (n=832)	Topical (n=653)	Geographic (n=79)	Corporate (n=40)	Topical-geographic (n=60)
Useful information	62.7	63.7	31.6	85.0	78.3
Too many retrievals	29.3	28.6	60.7	5.0	11.7
Too few retrievals	3.2	2.6	2.5	7.5	8.3
Off-the-mark	2.4	2.7	1.3	0.0	1.7
Useful NTs	1.1	1.1	1.3	2.5	0.0
Query too broad or narrow	0.9	1.0	1.3	0.0	0.0
Restate query	0.4	0.3	1.3	0.0	0.0
Total	100.0	100.0	100.0	100.0	100.0

Overall, almost two-thirds of exact matches retrieved useful information. Large percentages of queries bearing topical elements or geographic elements resulted in too many retrievals. This was especially true for geographic queries in which over sixty percent of retrievals were too large. Large retrievals were not a problem for matching corporate-body subject headings or for matching subject headings bearing topical and geographic elements. In fact, too few retrievals were as much of a problem for these two types of subject headings as too many retrievals. With regard to subject headings bearing topical and geographic elements, the matching topical *or* geographic element would have retrieved too many titles; the combination of these two different elements in a single heading reduced retrievals quite a bit.

A handful of matching subject headings or cross references were entirely off-the-mark with respect to the topics users had a mind. Table 4.9 explains user queries and off-the-mark matching terms.

Table 4.9. Off-the-mark Matching Terms

User query	Explanation
indian music	User is interested in “Music — India,” not the matching reference, “Indian music (American Indian).”
frog	User is interested in “Frog raising” which is a *see* reference to “Frogs — Culture.” The matching term is entirely off-the-mark, i.e., “FROG (Computer program language).”

english	User is interested in "Women," not matching reference, e.g., "English," which is a *see* reference to "British."
hudsons	User is interested in "Hudson Bay Co.," not matching name "Hudson's (Department store)."
conservation	User is interested in the heading "Wildlife conservation," not the matching heading "Conservation (Psychology)."
maya	User is interested in the heading "Mayas," not the matching heading "Maya (Hinduism)."

For a few queries — "indian music," "frog," "conservation," and "maya" — the existing search-tree configuration would be successful in retrieving titles on the topics that interest users but only if users continually pursue options to further their searches through subject searching approaches that follow the exact approach. For example, alphabetical searches would lead users rather quickly to the desired subject headings for the queries "frog" and "maya." Title-keyword searches would retrieve titles on "indian music" and "conservation" but users might have to view many less-than-useful titles to find ones that interest them — especially for the query "conservation" which is likely to retrieve this word in many different and undesired contexts.

The existing search-tree configuration would not be successful in retrieving titles on the "hudsons" query because the first word in the matching name heading, i.e., "Hudson Bay Co.," does not bear a trailing "s." If the one-word search tree included alphabetical searches as a last resort, this user might find the desired subject heading; however, he would have to review the results of several subject searches before the alphabetical approach and it is doubtful he would persevere.

Table 4.10 describes queries too broad or narrow to express users' topics of interest.

Table 4.10. Queries Too Broad or Narrow to Express Users' Topics of Interest

User Query	**Explanation**
goals	User is interested in the subject "Goal setting in personnel management," not the matching heading, "Goals (Sports)."
artic [sic]	User is interested in the subject "Arctic wolves," not matching headings and references, e.g., "Arctic regions," "Arctic ground squirrel," "Arctic Ocean," "Arctic mythology."
africa	User is interested in "Africa — Economic conditions," not the broad matching heading, i.e., "Africa."
buildings	Users is interested in "forts of the 1800s" but matches the broad subject heading "Buildings" under which are no useful references to subject headings about forts. The term "Forts" is a *see* reference for "Fortification."

On occasion, the exact approach would deliver users to narrower terms or subdivided forms of the matched subject heading that expresses the topics they have in mind but have not made explicit in their queries. An example of the former is the search for "brothers and sisters;" the narrower term "Sibling rivalry" under this subject heading describes exactly what the user had in mind. An example of the latter is the search for "africa;" the subdivision "Economic conditions" describes exactly what the user had in mind but did not make explicit in her query.

Infrequently, users entered queries that were not close to the topics they had in mind. An example is the query "english." Subsequent queries demonstrate that this user is really interested in "Women." The only way to retrieve titles on this topic would be for the user to reenter his query using this term.

4.4Queries for the Alphabetical Approach

4.4.1 Introduction

Next, systems governed by search trees submitted queries that failed the criteria for the exact approach to the alphabetical approach (figures 3.3 and 3.4). Queries that were candidates for the alphabetical approach matched longer subject headings or cross references. Table 4.11 is a fact sheet for user queries for subjects generally that systems governed by search trees would submit to the alphabetical approach.

Table 4.11. Fact Sheet for Queries for the Alphabetical Approach

Facts	Total	Topical	Geog.	Corp.	Topical-geog.
Searches					
No. of searches	155	130	7	10	8
No. of access points	554	463	16	26	49
Avg. no. of access points per search	3.6	3.6	2.3	2.6	6.1
Maximum no. of access pts. in a search	26	26	5	5	19
Initial access points					
Avg. no. of words per access point	1.3	1.3	1	1.6	1.6
Maximum no. of words in an access pt.	4	4	1	3	3
Percentage of access points with zero retrievals	19.4*	16.9*	80.0*	28.6*	0.0*
Percentage of access points with retrievals > 999	7.2†	6.0†	0.0†	14.3†	20.0†

*Does not include LS/2000

†Does not include LS/2000 and MIRLYN.

Partial matches accounted for 155 queries. This number was 8.1% of the total number of queries in the study. The majority (83.9%) of partial matches were queries for topical subjects. Partial matches figured into searches that averaged between two and six access points per search. One partial match initiated a search that featured 26 access points. Partial matches averaged between one and two words per query. Queries did not exceed four words long; most were one word long. Examples of long and short queries are:

- gifted and talented education
- egyptian
- welfare
- food industry
- naval

Almost 20% of partial matches failed to retrieve bibliographic records in the catalog the user originally searched. Of the four catalogs in this study, only LS/2000 and MIRLYN featured the alphabetical search; however, we did not compare the results of queries in the partial match category that users entered using the alphabetical search with the results of such queries for other searches. High-posted searches were not very common.

4.4.2 Types of Alphabetical Matches

Search trees for subjects generally tested queries to determine if they were candidates for the alphabetical approach (figures 3.3 and 3.4). Thus, our analysis of user queries for subjects generally continued with this test. Table 4.12 shows the result. Of the total of 1,465 queries for subjects generally, 155 queries met the criteria for the alphabetical approach. These queries were partial matches in that they matched longer subject headings or references.

Table 4.12. Types of Partial Matches

Type of partial match	Total (N=155)	Topical (N=130)	Geographic (N=7)	Corporate (N=10)	Topical-geographic (N=8)
Two or more words in heading	20.0	16.2	14.3	50.0	50.0
Two or more words in reference	4.5	4.6	0.0	10.0	0.0
Two or more words in reference, spelling error(s)	1.9	1.5	0.0	0.0	12.5
One word in heading	23.2	21.5	42.9	40.0	12.5
One word in reference	23.9	26.2	14.3	0	25.0
One word in heading, spelling error(s)	2.6	1.5	28.5	0.0	0.0
Less than one word in heading	18.7	22.3	0.0	0.0	0.0

Less than one word in reference	4.5	5.4	0.0	0.0	0.0
Less than one word in heading, spelling error(s)	0.7	0.8	0.0	0.0	0.0
Total	100.0	100.0	100.0	100.0	100.0

Except for queries bearing topical elements only, partial matches for alphabetical searches were pretty rare. Table 4.13 lists examples of queries in several partial-match categories.

Table 4.13. Partial Matches for the Alphabetical Approach

Query	Matching heading or reference	Partial-match type
magazine	Magazines	Less than one word in reference
food industry	Food industry and trade	Two or more words in heading
training	Training, Circuit; Training, Cross-cultural; Training, Football; etc.	One word in reference
bride	Bride price	One word in heading
plant growth	Plant growth inhibiting substances	Two or more words in heading
dwi	Dwingelderveld (Netherlands)	Less than one word in heading
carribean	Caribbean literature (French)	One word in heading, spelling error
chernoyble	Chernobyl Nuclear Accident , Chernobyl', Ukraine	One word in heading, spelling error
crcreative	Creative ability	One word in heading, spelling error
oorrientalism	Orientalism in art	One word in reference, spelling error

4.4.3 Match Satisfaction

Partial matches did not guarantee that the matched headings and/or references would guide users to useful information on their topics of interest. Figure 4.1 categorizes types of end-user queries by satisfaction categories.

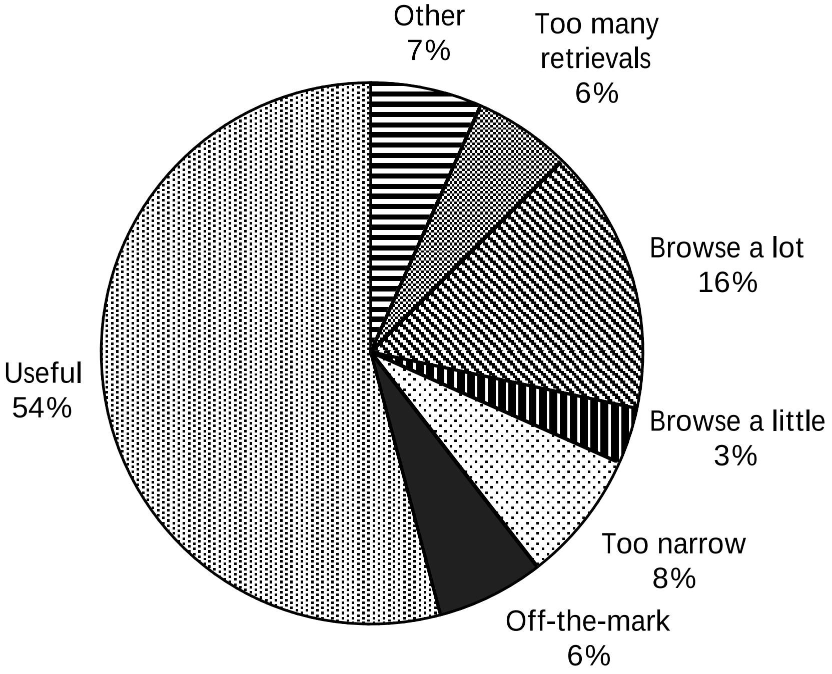


Figure 4.1. Satisfaction categories for partial matches

Over half of matching subject headings and references would be useful to users in terms of retrieving titles on their topics of interest. Examples of partial matches that have great potential for guiding users to subject headings that describe their topics of interest are:

User query	Matching heading or reference
kent state	Kent State University
coca-cola	Coca-cola Bottling Co.
congress	Congress facilities; Congresses and conventions; etc.
march on washington	March on Washington for Jobs and Freedom, Washington, D. C., 1963
documentary	Documentary bills; Documentary credit; Documentary drama; etc.
strain	Strain gages; Strain hardening; Strains and stresses; etc.
cad/cam	CAD/CAM systems
dinasour (sic)	Dinosaurs

quaker	Quaker bonnet (Plant); Quaker church buildings; etc.
circuit analysis	Circuit analysis, Electrical

A few partial matches, e.g., “Circuit analysis, Electrical,” “March of Washington…,” “CAD/CAM systems,” described topics that might have satisfied users. On occasion, the closest match did not describe users’ queries but a suitable term(s) would be listed on the same screen as the closest partial match, e.g., “Documentary drama” and “Strains and stresses.”

Browsing more than an initial screen of alphabetically-arranged subject headings would be required for almost 20% of user queries in the partial-match category. Table 4.14 explains user queries and partially-matching terms that users would find through considerable browsing.

Table 4.14. Browsing for Suitable Terms

User Query	Explanation
aurie	User is interested in “Arles (France).” She would have to browse backward many times to find this place.
naval	User is interested in scuttling of warships. He would have to browse forward to find reference “Naval ships,” for “Warships.” A subdivided form of this heading might produce useful titles.
afroamerican	User would have to browse the many subject headings beginning with “Afro-American” and focus her search.
instruments	User is interested in “Instruments, Musical” and would have to browse the many subject headings beginning with “instruments” to find this reference.
grocery	User would have to browse the many subject headings beginning with “grocery” and focus his search.

Matching headings and references placed in the “too narrow” satisfaction category did not adequately describe users’ topics because they were too narrow in meaning. Into the “off-the-mark” satisfaction category were placed matching headings and references that described topics different from the topics users had in mind. Table 4.15 gives examples of both satisfaction categories.

Table 4.15. “Off-the-mark” and “Too Narrow” Categories (Partial Matches)

User Query	Explanation
training	User is interested in training for using computers. There are many references beginning with the term “Training” but none are on computers. The user should add the term “computers” to this query.

bride	Nearby terms, e.g., “Bride price,” “Bridal bouquets,” and “Bridal gowns,” do not describe the topic the user has in mind (“Weddings”). A title-keyword search produces useful titles.
job	Nearby terms, e.g., “Job analysis,” “Job applications,” “Job descriptions,” do not describe the topic the user has in mind (“getting a job overseas”). Browsing might yield a useful narrower term under “Job hunting” or a subdivided form of this heading.
fortifications	Matching reference “Fortifications, Attack and defence of” is too narrow. The user’s best bet is to browse backward to “Fortification” but he must persevere to reach this term.
woodworking	Matching heading “Woodworking industries” is off-the-mark in view of the user’s interest in “Cabinet-work.” Browsing backward to “Woodwork” and scanning related terms for ”Cabinet-work” might help out this user.
aircraft	Matching reference “Aircraft, Antisubmarine” is off-the-mark in view of the user’s interest in “Airplanes.” Browsing forward to “Aircraft industry” and scanning related term ”Airplanes” might help out this user.
media	Matching headings and references, e.g., “Media, Potting,” “Media programs (Education),” “ Media selection,” are off-the-mark in view of the user’s interest in “Mass media.” Following up this search with a keyword-in-main-heading search might produce useful titles.

There is not a single way of getting users back on track following the retrieval of terms that are off-the-mark or too narrow for their interests. On occasion, subsequent subject searches chosen by the search trees will yield useful titles. For example, title-keyword and keyword-in-main-heading searches would have gotten the users who entered queries for “bride” and “media,” respectively, back on track. Some users must browse backward or forward, choose a listed subject heading, and pursue broader, related, or narrower terms connected with the selected heading. Other users might have to restate their queries to retrieve useful titles.

The problem of too many retrievals occurred infrequently in connection with partial matches. Yet, a few queries, e.g., “italian,” “food industry,” “magazine” could have resulted in many retrievals because of the large number of subject headings that began with these terms.

4.5Queries for Keyword-in-heading Searches

4.5.1 Introduction

Next, systems governed by search trees submitted queries that failed the criteria for the exact and alphabetical approaches to keyword-in-main-heading and keyword-in-subdivided-heading searches (figures 3.5A). Queries that were candidates for keyword-in-heading searches must have contained the same words as unsubdivided subject headings (for the keyword-in-main-heading search) or subdivided subject headings (for the keyword-in-subdivided-heading search). Table 4.16 is a fact sheet for user queries for subjects generally that systems governed by search trees would submit to keyword-in-heading searches.

Table 4.16. Fact Sheet for Queries for Keyword-in-heading Searches

Facts	Total	Topical	Geog.	Corp.	Topical-geog.
Searches					
No. of searches	98	53	3	0	42
No. of access points	348	201	11	0	136
Avg. no. of access points per search	3.6	3.8	3.7	0	3.2
Maximum no. of access pts. in a search	19	19	6	0	16
Initial access points					
Avg. no. of words per access point	2.4	2.2	2.3	0	2.6
Maximum no. of words in an access pt.	8	4	3	0	8
Percentage of access points with zero retrievals	20.2*	17.8*	0.0*	0	23.8*
Percentage of access points with retrievals > 999	0.0†	0.0†	0.0†	0	0.0†

*Does not include LS/2000
†Does not include LS/2000 and MIRLYN.

Keyword-in-heading matches accounted for 98 queries. This number was 5.1% of the total number of queries in the study. The majority (54.1%) of keyword-in-heading matches were queries for topical subjects; however, queries bearing elements for topics and geographic names also contributed a large share (42.9%) of keyword-in-heading matches. No queries for corporate names were keyword-in-heading matches. Keyword-in-heading matches figured into searches that averaged between three and four access points per search. Keyword-in-heading matches averaged about two words per query. This average was higher than averages for exact (1.6 words, Table 4.5) and partial (1.3 words, Table 4.11) matches. Examples of keyword-in-heading matches are:

- aids advertising
- italian linguistics
- economy africa
- automatic processing
- retail business

About 20% of keyword-in-heading matches of topical subjects and geographic names failed to retrieve bibliographic records in the catalog the user originally searched. High-posted searches were not characteristic of keyword-in-heading matches.

4.5.2 Types of Keyword-in-heading Matches

Of the total of 1,919 queries in the study, 98 (5.1%) queries met the criteria for the keyword-in-heading searches. These queries were partial matches in that the subject headings they matched contained additional words besides the matched words; however, all words in users queries had to occur in matched subject headings. Table 4.17 details the specifics on types of keyword-in-heading matches.

Table 4.17. Types of Keyword-in-heading Matches

Type of partial match	Total (N=98)	Topical (N=53)	Geographic (N=3)	Corporate (N=0)	Topical-geographic (N=42)
Main heading	24.5	34.0	0.0	0	14.3
Subdivided heading	66.3	62.2	100.0	0	69.0
Subdivided heading, spelling error	9.2	3.8	0.0	0	16.7
Total	100.0	100.0	100.0	0	100.0

The majority of keyword-in-heading matches were matches of subdivided subject headings. Only about one-quarter of matches were matches of main, unsubdivided subject headings.

4.5.3 Match Satisfaction

Keyword-in-heading matches did not guarantee that the matched headings and/or references would guide users to useful information on their topics of interest. Figure 4.2 categorizes types of end-user queries by satisfaction categories.

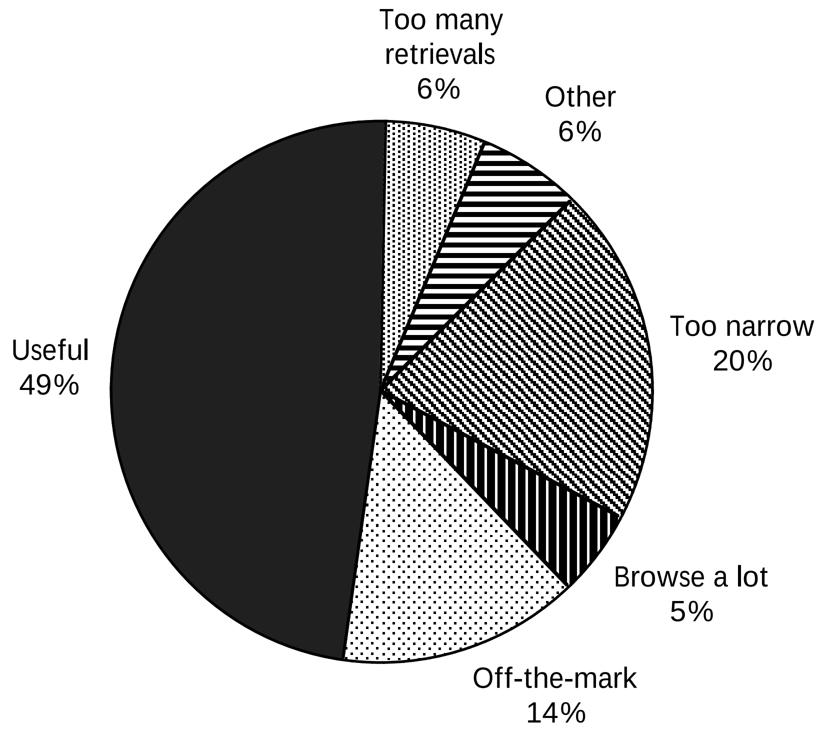


Figure 4.2. Satisfaction categories for keyword-in-heading matches

A little under half of matching subject headings were useful in terms of addressing the topics of user queries. Examples of matching subdivided and unsubdivided subject headings are:

User query	Matching subject headings
maps salina	Salina (Kans.) — Maps
israel conflict	Israel-Arab conflicts
south africa and te (sic) church	South Africa — Church history
singapore health	Health planning — Singapore; Health facilities — Singapore; etc.
paris architecture	Architecture — France — Paris
communication international aspects	Communication, International, subdivided by several subdivisions bearing the word "aspects"
japanese phonetics	Japanese language — Phonetics
women status in africa	Women — Legal status, laws, etc. — Africa

Matching subject headings that were off-the-mark or too narrow to satisfy users' interests were major problems with keyword-in-heading matches. A little more than a third of keyword-in-heading matches were plagued by these problems. Table 4.18 describes a few of these problems and possible solutions.

Table 4.18. "Off-the-mark" and Too Narrow" Categories (Keyword-in-heading Matches)

User Query	**Explanation**
instruction women	User is interested in "Women — Education, Medieval." The heading matching the user query, "Sex instruction for women," is off-the-mark. The user's best bet is to restate his query.
love and relationships	The matching heading "Love-hate relationships" is too narrow to express the user's interests. Title-keyword searches yield potentially useful titles such as *Life & love, such as they are*, *At long last love: sage advice and true stories from America's premier matchmakers,* and *Lasting love relationships.*
stress college students	The matching heading "College students — Job stress" is too narrow to express the user's interests. Keyword searches of subject heading fields yield several potentially useful titles bearing the subject headings "College students" and "Stress (Psychology)," e.g., *Stress in college students, A textbook of stress for college students*, and *Student stress: effects and solutions.*
retail business	The matching heading "Business mathematics — Retail trade" is off-the-mark. Keyword searches of subject heading fields yield several potentially useful titles bearing the subject heading "Retail trade" and several subjects headings bearing the word "business," e.g., "Central business districts," "Small business," and "Business enterprises."
economy europe	The matching heading "Mixed economy — Europe" is too narrow for the user's interest. Title-keyword searches retrieve several potentially useful titles, e.g., *The economic emergence of a new Europe, Singular Europe: economy and politics of the European Community after 1992,* and *Towards a new Europe?: structural change in the European economy.*
2slave (sic) religion	The matching heading "Slave Indians — Mythology and religion" is too narrow for the user's interest. Keyword-in-record searches retrieve several potentially useful titles, e.g., *Dark symbols, obscure signs: God, self, and community in the slave mind, Cut loose your stammering tongue: Black theology in the slave narratives,* and *Slavery and the slave-holder's religion.*

refugee resettlement	The matching heading “U. S. Office of Refugee Resettlement” may be too narrow to satisfy the user’s interests. Keyword-in-record searches yield several potentially useful titles, e.g., *Refugee resettlement and wellbeing*, *Refugee mental health in resettlement countries*, and *Vietnamese in America: an analysis of adaptational patterns*.
indian women	The matching reference “Women, Indian” directs users to a subject heading, “Indians of North America — Women,” that is off-the-mark. Title-keyword searches yield the following potentially useful titles: *Indian women: challenges and change*, *Indian women: images and reflections*, *Indian women’s movement*.

Systems governed by search trees would retrieve potentially useful titles for most of the queries listed in Table 4.18. They would retrieve them in subject searches following keyword-in-heading searches, viz. title-keyword, keyword searches of subject heading fields, and keyword-in-record searches. Of course, users would have to persevere and review the results of one or more follow-up searches.

“Too many retrievals” and “lots of browsing” characterized about 10% of keyword-in-heading matches. Examples of queries placed in these satisfaction categories are:

User query	Matching subject headings
southern literature	“Southern states in literature,” and many “literature ”headings subdivided by “Southern states,” e.g., “American literature — Southern states,” “Women and literature — Southern states.”
cultural studies	Many subject headings subdivided by “Cross-cultural studies”
soviet union military	Many subject headings with these words in main headings and subdivisions, e.g., “Military scouts — Soviet Union,” “Military doctrine — Soviet Union,” “Soviet Union — Military policy,” “Soviet Union — Military relations.”
world war 1939 1945 aerial	Many subject headings beginning with “World War, 1939–1945 — Aerial operations” that are geographically subdivided.
law california	Many subject headings bearing the subdivisions “California” and “Law and legislation.”

User perseverance was the key to the searches described above. Users might have to browse dozens, even hundreds of subject headings to find the one(s) that describe their topics of interest. For example, the query “soviet union military” retrieves many dozens of subject headings beginning with “Soviet Union” in which “military” occurs in subdivisions and vice versa. Some users might not have enough patience to scan long lists of subject headings, and, instead, terminate their search or enter new queries.

"Few retrievals" characterize a few keyword-in-heading matches. Examples are the queries "australia retail" and "tax forms;" these queries matched the subject headings "Retail trade — Australia" and "Income tax — Forms," respectively, but these headings retrieved few titles. In systems governed by search trees, keyword-in-record searches could retrieve additional titles.

4.6Queries for Keyword Searches

4.6.1 Introduction

Lastly, systems governed by search trees submitted queries that failed the criteria for exact, alphabetical, and keyword-in-heading searches to keyword searches (figure 3.5B). To yield retrievals, queries must have contained the same words as titles (for title-keyword searches), subject heading fields of bibliographic records (for keyword searches of subject heading fields), or subject-bearing fields of bibliographic records (for keyword-in-record searches). Table 4.19 is a fact sheet for user queries for subjects generally that systems governed by search trees would submit to keyword searches.

Table 4.19. Fact Sheet for Queries for Keyword Searches

Facts	Total	Topical	Geog.	Corp.	Topical-geog.
Searches					
No. of searches	290	249	3	1	37
No. of access points	934	796	10	8	120
Avg. no. of access points per search	3.2	3.2	3.3	8	3.2
Maximum no. of access pts. in a search	36	36	4	8	9
Initial access points					
Avg. no. of words per access point	2.5	2.4	2.3	2	2.9
Maximum no. of words in an access pt.	6	6	3	2	6
Percentage of access points with zero retrievals	50.2*	48.0*	100.0*	100.0*	59.5*
Percentage of access points with retrievals > 999	0.0†	0.0†	0.0†	0.0†	0.0†

*Does not include LS/2000
†Does not include LS/2000 and MIRLYN.

Keyword matches accounted for 290 queries. This number was 15.1% of the total number of queries in the study. The majority (85.9%) of keyword matches were queries for topical subjects. Keyword matches initiated searches that averaged over three access points per search. One keyword match initiated a search that featured 36 access points; this search began with the access point “latinas” and included many other queries on feminism, literature, and Hispanic women, e.g., “american literature — hispanic american authors,” “feminism and hispanic americans,” “feminism and latina women,” “feminism,” “sex role and hispanic americans,” and “feminism — hispanic american authors.” Keyword matches averaged between two and three words per query.

Keyword matches often failed to retrieve bibliographic records in the catalog the user originally searched. Overall, about half of user queries failed to retrieve titles. There were no high-posted searches.

4.6.2 Types of Keyword Matches

Of the total of 1,919 queries in the study, 290 queries (15.1%) were successful in retrieving bibliographic records through various keyword searches. Table 4.20 details specifics about keyword matches.

Table 4.20. Types of Keyword Matches

Type of partial match	Total (n=290)	Topical (n=249)	Geog. (n=3)	Corporate (n=1)	Topical-geographic (n=37)
Two or more title words	73.5	73.9	100.0	100.0	67.6
Two or more title words, spelling error	4.1	4.8	0.0	0.0	0.0
One title word	7.2	8.0	0.0	0.0	2.7
One title word, spelling error	0.7	0.8	0.0	0.0	0.0
Words in subject heading fields	3.8	2.8	0.0	0.0	10.8
Words in subject-rich fields	10.0	8.9	0.0	0.0	18.9
Words in subject-rich fields, spelling error	0.7	0.8	0.0	0.0	0.0
Total	100.0	100.0	100.0	100.0	100.0

Over 85% of keyword matches were title matches. Such matches would provoke systems governed by search trees to respond to matches with title-keyword searches. Less than 4% of keyword matches would provoke systems to respond with keyword searches of subject heading fields. A little over 10% of keyword matches would result in keyword-in-record searches.

4.6.3 Match Satisfaction

Figure 4.3 categorizes keyword matches of end-user queries by satisfaction categories.

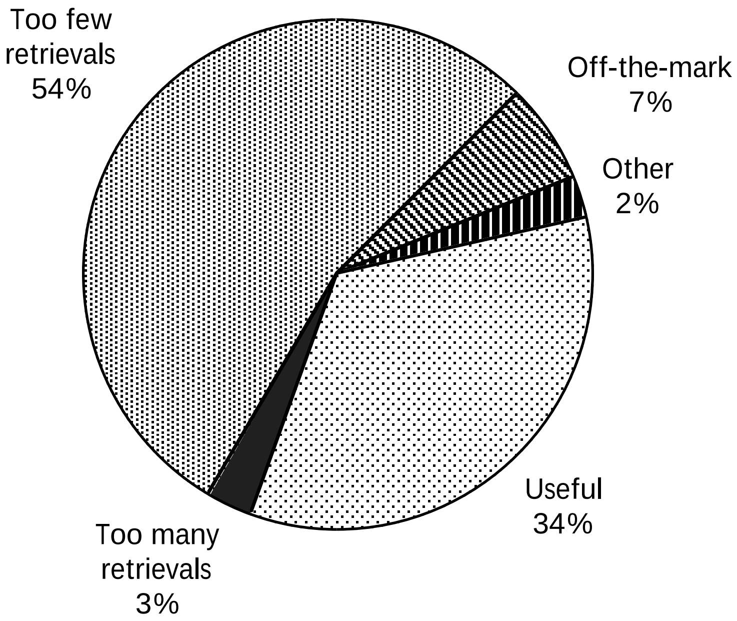


Figure 4.3. Satisfaction categories for keyword matches

For the first time in this analysis, the “useful” satisfaction category did not account for a majority of queries. Instead, the “too few retrievals” category accounted for the majority (54%) of queries. Since keyword searches were the last searches that the search trees considered in the existing configuration, enhancements to this configuration should include new strategies for finding additional titles on users’ topics of interest. Table 4.21 describes keyword searches for queries that retrieved too few titles and strategies for finding additional ones.

Table 4.21. Strategies for Enhancing “Too Few Retrievals”

User Query	Explanation
tournaments medieval	Keyword-in-record searches retrieve less than five titles in Michigan’s MIRLYN online catalog: *The medieval tournament, City, marriage, tournament: arts of rule in late medieval Scotland, The tournament in England, 1100–1400.* The subject heading “Tournaments — History” occurs in two titles and could be used to find additional titles.

1960s riots	Title-keyword searches in MIRLYN yield two titles: *Black violence: political impact of the 1960s* and *A study of arrest patterns in the 1960's riots.* Occurring in several titles are "Riots — United States," "Afro-Americans — History," and "Passive resistance" which could be used to find additional titles.
high fidelity stereo	Keyword-in-record search yields one title (*The stereo high fidelity handbook)* with the subject heading "Stereophonic sound systems" which could be used to find additional titles.
maya wars	Keyword-in-record search yields three titles. Two of the titles bear the subject heading "Mayas — Wars" which could be used to find additional titles.
military and gays	Keyword-in-record searches yield eight titles. Examples are *Enlisted meat, and other true military homosexual stories, Conduct unbecoming: lesbians and gays in the U. S. military, Vietnam to the Persian Gulf, Gays in uniform,* and *Fighting back: lesbian and gay draft, military, and veterans issues.* Most are assigned the subject heading "United States — Armed forces — Gays" which could be used to find additional titles.
feminism and race	Keyword searches of subject heading fields yield a dozen titles. Examples are *Segregated sisterhood: racism and the politics of American feminism, Daughters of Jefferson, daughters of bootblacks: racism and American feminism,* and *Common differences: conflicts in Black and White feminist perspectives.* Common subject headings are "Feminism — United States," "United States — Race relations," and "Racism — United States" which could be used to find additional titles.
stereoscopy	Keyword-in-record searches yield two titles. One of the two titles bears the subject heading "Photography, Stereoscopic" which could be used to find additional titles.

Keyword searches for the queries listed in Table 4.21 yielded a dozen or fewer titles. On occasion, a single subject heading was assigned to several retrieved titles (see explanations for searches on "maya and wars," "military and gays," and "tournaments medieval"). Sometimes, two or more subject headings were assigned to several retrieved titles (see explanations for searches on "feminism and race" and "1960s riots"). Other queries retrieved so few titles that a single subject heading might be the only information in retrieved records that could be used to further the search (see explanations for searches on "stereoscopy" and "high fidelity stereo").

For a few queries, keyword matches still produced too many retrievals. Examples are the queries "social services" and "tourism." Systems could help users focus their searches by displaying subject headings common to many retrievals and using related terms and subdivided forms of common subject headings as devices for refining searches. For example, the query "tourism" produced many titles with the subject headings "Tourist trade" and "Tourist trade and state." Systems could suggest to users that they focus and refine their searches using subdivided forms of these subject headings.

Queries that were "off-the-mark" amounted to 7% of keyword matches. Included in the "other" category were a few queries for which retrievals were "too narrow" (1.4%). Table 4.22 discusses queries that resulted in "off-the-mark" and "too narrow" retrievals.

Table 4.22. "Too Narrow" or "Off-the-mark" Categories (Keyword Matches)

User Query	Explanation
future prices commodities	Keyword-in-record search yields titles such as *Natural resource commodities* and *Trends in natural resource commodities.*
saw horse	Keyword-in-record search yields *The marvelous land of Oz: being an account of the further adventures of the Scarecrow and Tin Woodman … the animated saw-horse …*
american ceramists	Keyword-in-record search yields *Phase diagrams for ceramists* which is published by the American Ceramic Society.
swiss universities	Keyword-in-record search yields *Dissertations in English and American literature: theses accepted by Austrian, French, and Swiss universities* and *Canon law in Protestant lands.*

Keyword-in-record searches for the four queries in Table 4.22 yielded false drops. Since keyword-in-record searches were the "search of last resort" in the existing search-tree configuration, the results of these keyword-in-record searches would not satisfy users. Finding additional titles would require systems to truncate query words, and, possibly, for queries exceeding two words, invoke the best-match approach. The best-match approach would feature stemming, weighted-term probabilistic retrieval, and output ranking. Systems would perform searches on a best match, combinatorial basis. A match on fewer than all query-word stems might retrieve titles but titles that have all the query-word stems would be displayed first. Query-word stems would be weighted so frequently-occurring stems would get a low "importance" weight compared to the weight assigned to rare word stems.

4.7Chapter Summary

Chapter 4 discusses the characteristics of four types of matches of online catalog subject vocabulary: (1) exact, (2) partial, (3) keyword-in-heading, and (4) keyword matches. Search trees can handle the four match types discussed in this chapter by submitting particular types to exact, alphabetical, keyword-in-heading, and keyword-in-record searches.

Exact matches accounted for 832 queries (section 4.3). This number was 43.4% of the total number of queries in the study (Table 4.5). Exact matches would be candidates for the exact search in bibliographic systems. The majority (78.5%) of exact matches were queries for topical subjects. Queries for geographic names came in a distant second in that they accounted for only 9.5% of exact matches. Queries for topical subjects and geographic names were likely to be exact matches of subject headings (Table 4.6). Queries for corporate names typically matched references. Queries bearing elements for both topical subjects and geographic names were likely to match normalized forms of subject headings or references.

Overall, almost two-thirds of exact matches retrieved useful information (Table 4.8). Large percentages of queries bearing topical elements or geographic elements resulted in too many retrievals. This was especially true for queries bearing geographic names in which over sixty percent of retrievals were too large. Large retrievals were not a problem for matching corporate-body subject headings or for matching subject headings bearing topical and geographic elements. In fact, too few retrievals were as much of a problem for these two types of subject headings as too many retrievals.

Partial matches accounted for 155 queries (section 4.4). This number was 8.1% of the total number of queries in the study (Table 4.11). Partial matches would be candidates for the alphabetical search in bibliographic systems. The majority (83.9%) of partial matches were queries for topical subjects. A little more than half of matching subject headings and references would be useful to users in terms of retrieving titles on their topics of interest (figure 4.1). Browsing more than an initial screen of alphabetically-arranged subject headings would be required for almost 20% of partially-matching user queries. Browsing many subject headings beginning with the same word(s) as user queries would require end users' patience and perseverance.

Keyword-in-heading matches accounted for 98 queries (section 4.5). This number was 5.1% of the total number of queries in the study (Table 4.16). Keyword-in-heading matches would be candidates for keyword-in-heading searches in bibliographic systems. The majority (54.1%) of exact matches were queries for topical subjects; however, queries bearing elements for topics and geographic names also contributed a large share (42.9%) of keyword-in-heading matches. Keyword-in-heading matches averaged about two words per query. This average was higher than average for exact (1.6 words, Table 4.5) and partial (1.3 words, Table 4.11) matches. A little under half of matching subject headings were useful in terms of addressing the topics of user queries (figure 4.2). Matching subject headings that were off-the-mark or too narrow to satisfy users' interests were major problems with keyword-in-heading matches (Table 4.18).

Keyword matches accounted for 290 queries (section 4.6). This number was 15.1% of the total number of queries in the study (Table 4.19). Keyword matches would be candidates for various keyword searches in bibliographic systems (i.e., title-keyword, keyword in subject heading fields, keyword-in-record searches). The majority (85.9%) of keyword matches were queries for topical subjects. The “too few retrievals” satisfaction category accounted for the majority (54%) of keyword matches (figure 4.3). Since keyword searches were the last searches that the search trees considered in the existing configuration, enhancements to this configuration should include new strategies for finding additional titles on users’ topics of interest. Table 4.21 describes keyword searches in which subject headings in retrieved titles could be used to further the search.

5 No Matches

5.1 Introduction

Of the 1,919 queries for subjects generally, 90 (4.7%) failed to meet the criteria for exact, alphabetical, keyword-in-heading, and keyword matches. This chapter takes an in-depth look at these queries, suggests ways in which the existing search-tree configuration could be enhanced to handle them, and tests the enhanced configuration using the non-matching queries in this study.

5.2 Basic Information

Table 5.1 is a fact sheet for user queries that failed to meet the criteria for exact, alphabetical, keyword-in-heading, and keyword matches.

Table 5.1. Fact Sheet for Queries for No Matches

Facts	Total	Topical	Geog.	Corp.	Topical-geog.
Searches					
No. of searches	90	71	0	8	11
No. of access points	356	308	0	18	30
Avg. no. of access points per search	4.0	4.3	0	2.3	2.7
Maximum no. of access pts. in a search	24	24	0	6	7
Initial access points					
Avg. no. of words per access point	3.2	3.1	0	2.8	4.5
Maximum no. of words in an access pt.	10	8	0	5	10
Percentage of access points with zero retrievals	100.0*	100.0*	0*	100.0*	100.0*
Percentage of access points with retrievals > 999	0.0†	0.0†	0†	0.0†	0.0†

*Does not include LS/2000

†Does not include LS/2000 and MIRLYN.

Of the total of 1,919 queries in the study, 90 queries failed to meet the criteria for invoking subject searches on the existing configuration of search trees. This number was only 4.7% of the total number of queries in the study. The majority (78.9%) of non-matches were queries for topical subjects. Non-matches figured into lengthy searches that averaged four access points per search. One exact match initiated a search that featured 24 access points; this search began with the access point "boxing tragedies" and included many other queries on sports deaths, e.g., "dying in sports," "deaths (boxing)," "deaths (sports)," "killing in boxing," and "sporting (deaths)." Non-matches averaged over three words per query. All non-matches failed to retrieve titles. Examples of non-matching queries are:

- jews in post war era
- anti social personality disorder
- glycolysis and respiration
- retibol
- geisha dancers

Figure 5.1 shows the elements in non-matching queries.

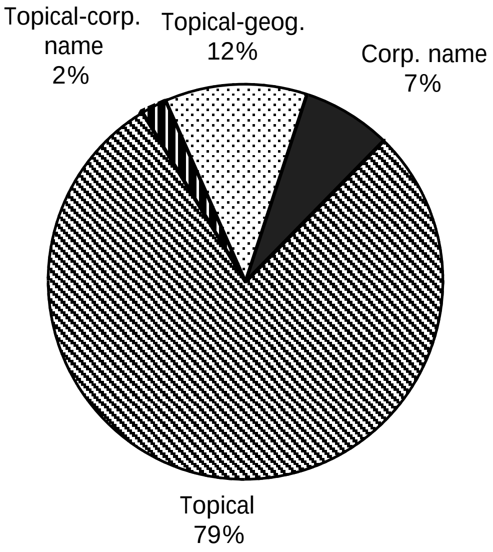


Figure 5.1. Elements in non-matching queries

The vast majority of non-matching queries were made up of topical elements. Coming in a distant second were queries bearing topical and geographic-name elements. There were no non-matching queries made up exclusively of geographic elements.

5.3Redesigned Search Trees

An analysis of online retrieval test data from a related study demonstrated that search trees were more effective in selecting a subject searching approach that would produce useful information for the subjects users seek than users would select on their own (Drabenstott and Weller 1994, chapter 13). The related study also demonstrated needed enhancements to the search-tree configuration to enable bibliographic systems to respond with useful retrievals to especially difficult user queries, that is, those for which the existing set of search trees produced no retrievals. Since the search trees tested in the related study were the same as the search trees used in this study to categorize end-user queries (see chapter 3), we will introduce the enhanced search trees from the related study in this section and use them to determine their effectiveness responding with useful retrievals to especially difficult user queries, that is, those for which the original search-tree configuration produced no retrievals in this study.

5.3.1 Initial Search Tree

The initial search tree remained basically unchanged from the original initial search tree (see figure 3.1). The only change from the original design was to return users whose searches failed to produce retrievals to the question about personal names. Based on user responses to this question in the related study, the new design of the initial search tree (figure 5.2) allowed users to distinguish their new or revised 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.

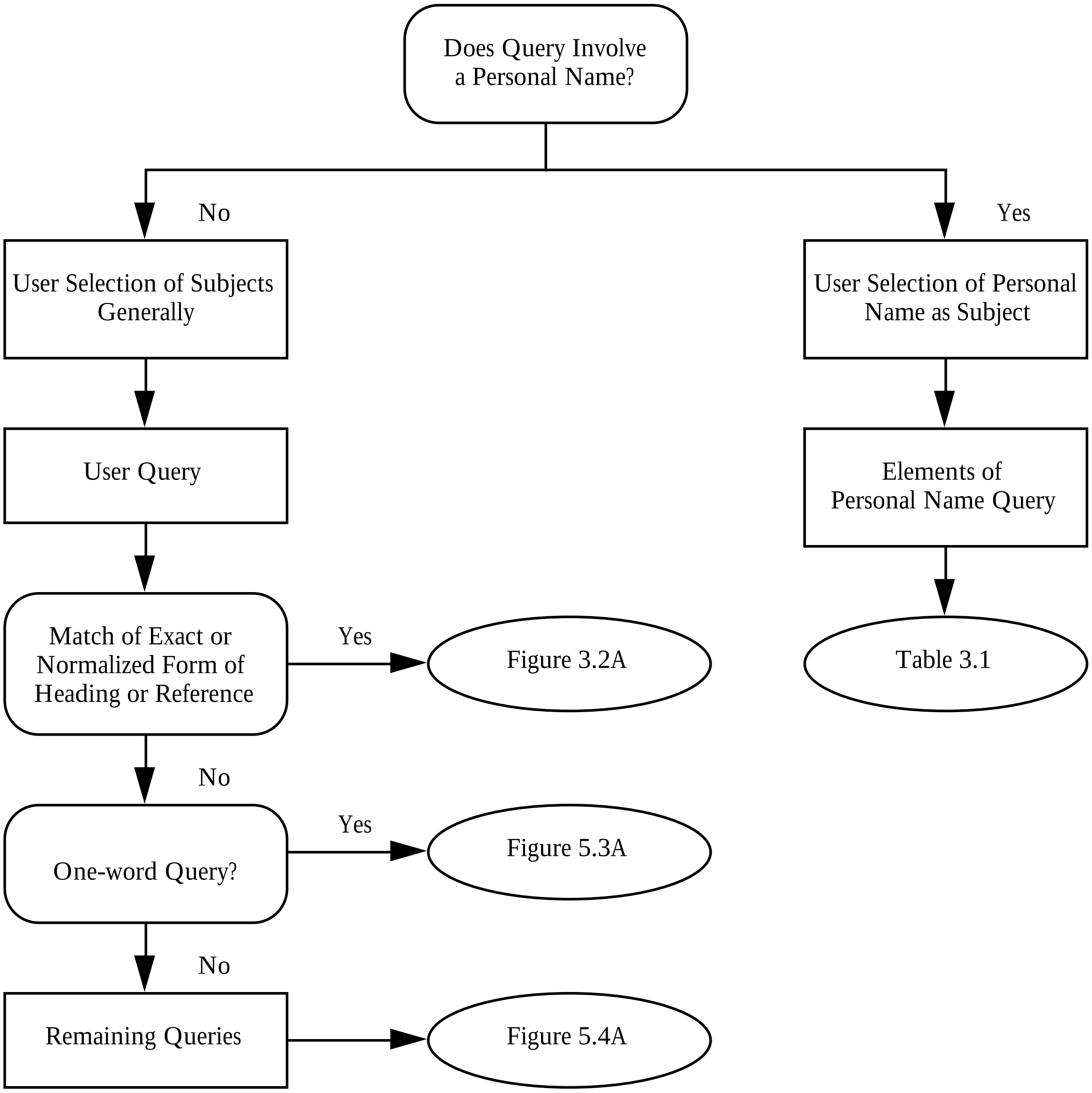

Does Query Involve a Personal Name?
No
Yes
User Selection of Subjects Generally
User Selection of Personal Name as Subject
User Query
Elements of Personal Name Query
Match of Exact or Normalized Form of Heading or Reference
Yes
Figure 3.2A
Table 3.1
No
One-word Query?
Yes
Figure 5.3A
No
Remaining Queries
Figure 5.4A

Figure 5.2. Redesigned initial search tree

The search tree for the exact approach remained unchanged. Systems responded to user queries for subjects generally that matched exact or normalized forms of controlled vocabulary terms with the exact approach. Figures 3.2A and 3.2B show the search tree that features the exact approach.

5.3.2 Search Tree for One-word Queries

The redesigned search tree for one-word queries contained a few changes from the original search tree for one-word queries: (1) checking go/see lists, (2) redirecting misspelled queries to the question on personal names, (3) invoking the keyword-in-record search, (4) invoking relevance feedback following title-keyword and keyword-in-record searches, and (5) invoking the alphabetical approach as the search type “of last resort.” The search tree for one-word queries is given in figures 5.3A and 5.3B.

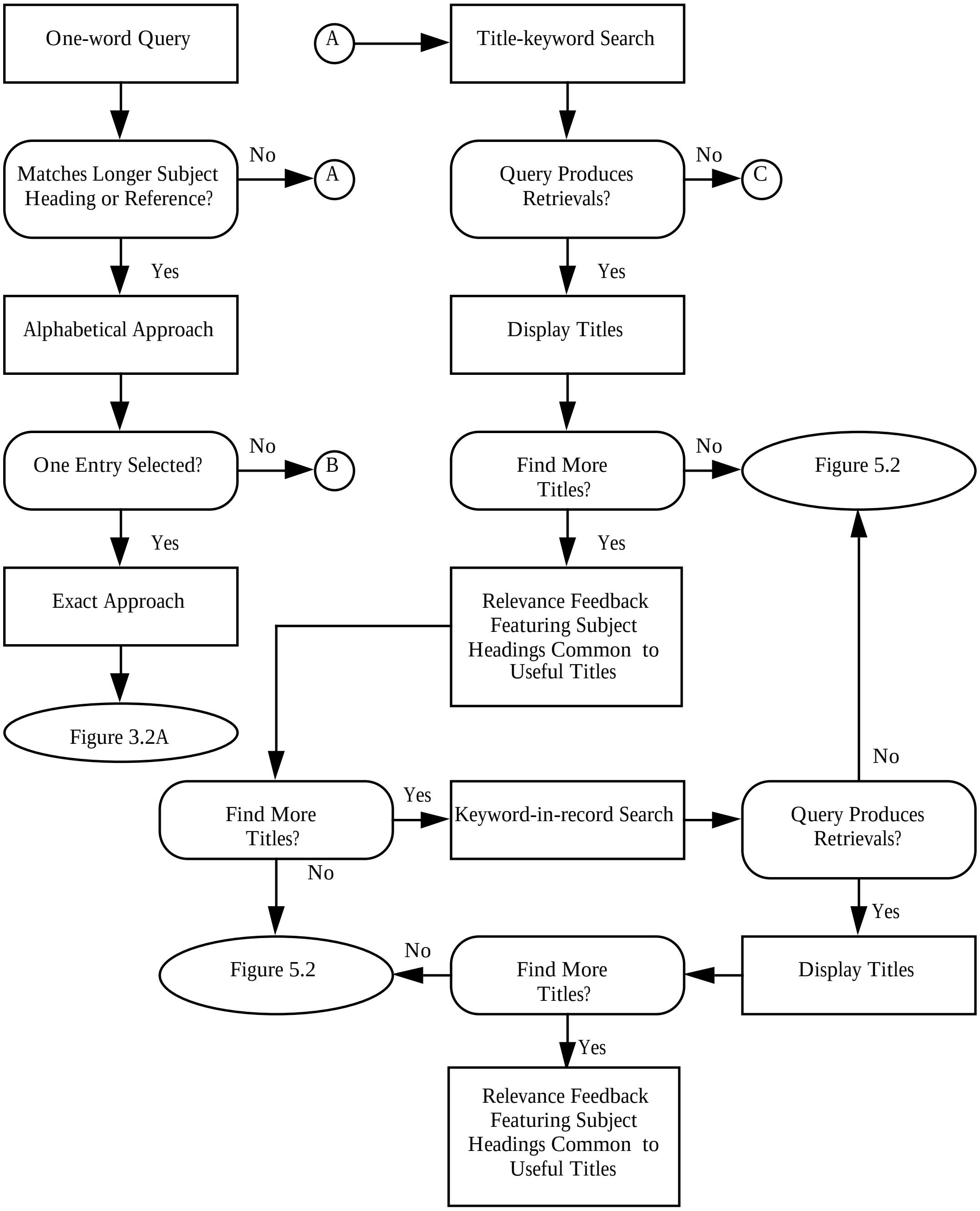


Figure 5.3A. Redesigned search tree for one-word queries

Figure 5.3A shows system actions that tested whether one-word queries met the criteria for invoking the alphabetical approach. Queries that failed were submitted to title-keyword searches. If they were successful, systems retrieved titles, and, at the conclusion of the title display, they asked users whether they wanted to retrieve additional titles. Systems responded to users who gave positive responses with the results of searches for subject headings common to several retrieved titles that users rated useful. Systems could continue to find additional titles through keyword-in-record searches and relevance feedback based on keyword-in-record search results. Of course, relevance feedback assumed that systems collected relevance assessments during the display of retrieved titles.

The original search tree for one-word queries did not include keyword-in-record searches (see figure 3.3). This search type was included in the enhanced version of the search tree for one-word queries. In view of the low levels of perseverance experimental online catalog users exhibited in the related study, few users will reach the results of keyword-in-record searches for one-word queries unless they continually expand search results. On occasion, however, keyword-in-record searches might be fruitful in retrieving useful titles for users whose queries failed to retrieve titles through exact, alphabetical, and title-keyword searches. For example, a contents note or summary field might bear a word matching the query. The only way for systems to retrieve this record would be through a keyword-in-record search.

Figure 5.3B shows the operations that systems performed on queries that failed to produce retrievals in title-keyword searches. They first checked user queries for possible misspellings. Systems could assist users in correcting spelling errors using spelling-correction routines similar to such routines for word processing programs in which they suggested alternate spellings for misspelled words. If users failed to correct spelling errors, systems could check queries against go/see lists to determine if queries contained listed words or phrases. They would then enhance queries with words and phrases from go/see lists and start at the initial search tree to find matches. Types of terms and phrases included in go/see lists would be one-word/two-word variations of concepts and noun/adjective pairs. If systems enhanced queries with words and phrases from go/see lists, search trees began with the initial search tree and tried to effect matches of the system-enhanced user query.

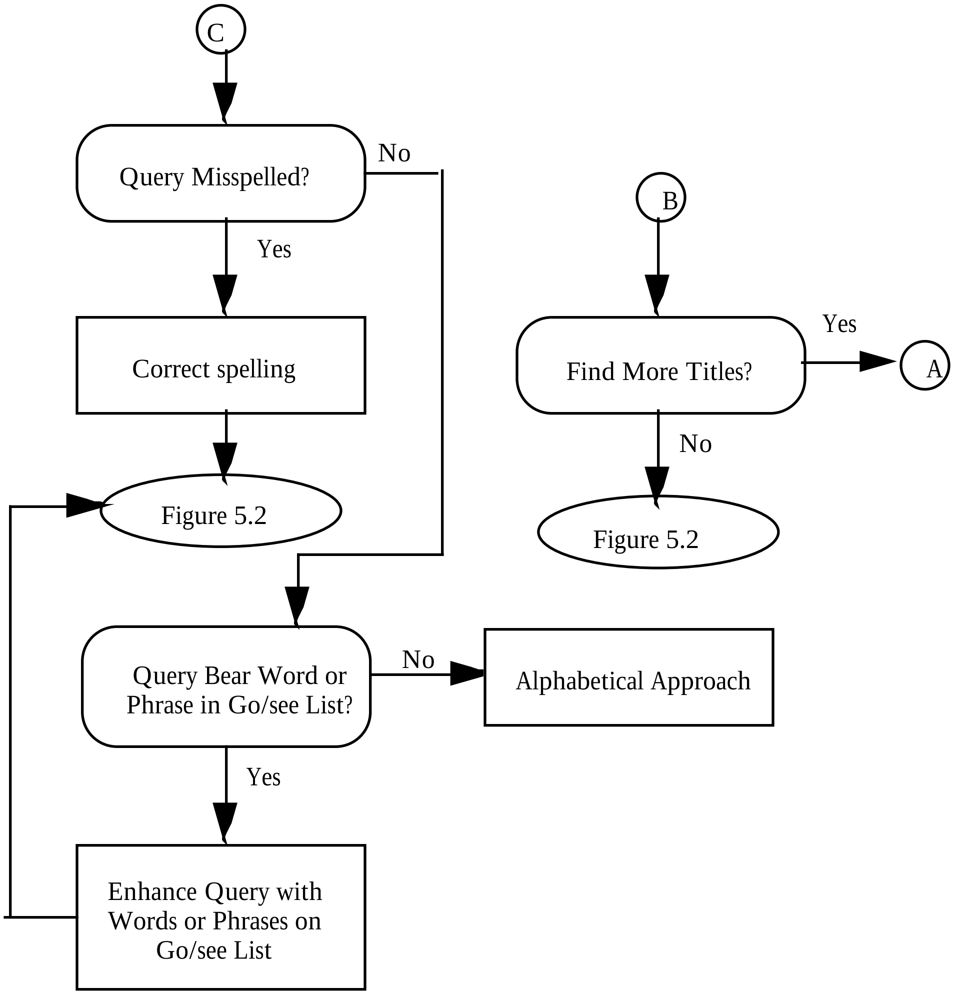


Figure 5.3B. Redesigned search tree for one-word queries (contd.)

The alphabetical approach was added to the search tree for one-word queries as the search type "of last resort." If all other subject searches failed to produce retrievals, systems would respond with the results of the alphabetical approach, i.e., a display of subject headings and *see* references in the alphabetical neighborhood of the user-entered term. If, however, the user query matched the criteria for an alphabetical search following the exact approach, the search type "of last resort" would be the keyword-in-record search.

5.3.3 Search Trees for Multi-word Queries

The redesigned search tree for multi-word queries featured the following enhancements: (1) checking go/see lists, (2) redirecting misspelled queries to the question on personal names, (3) performing relevance feedback based on useful retrievals in free-text searches (i.e., title-keyword, keyword in subject heading fields, and keyword-in-record searches), (4) invoking stemming, (5) invoking the best-match approach, and (6) invoking the alphabetical approach as the search type "of last resort." The redesigned tree is shown in figures 5.4A–C. The search tree for multi-word queries began with system efforts to invoke the alphabetical approach (figure 5.4A).

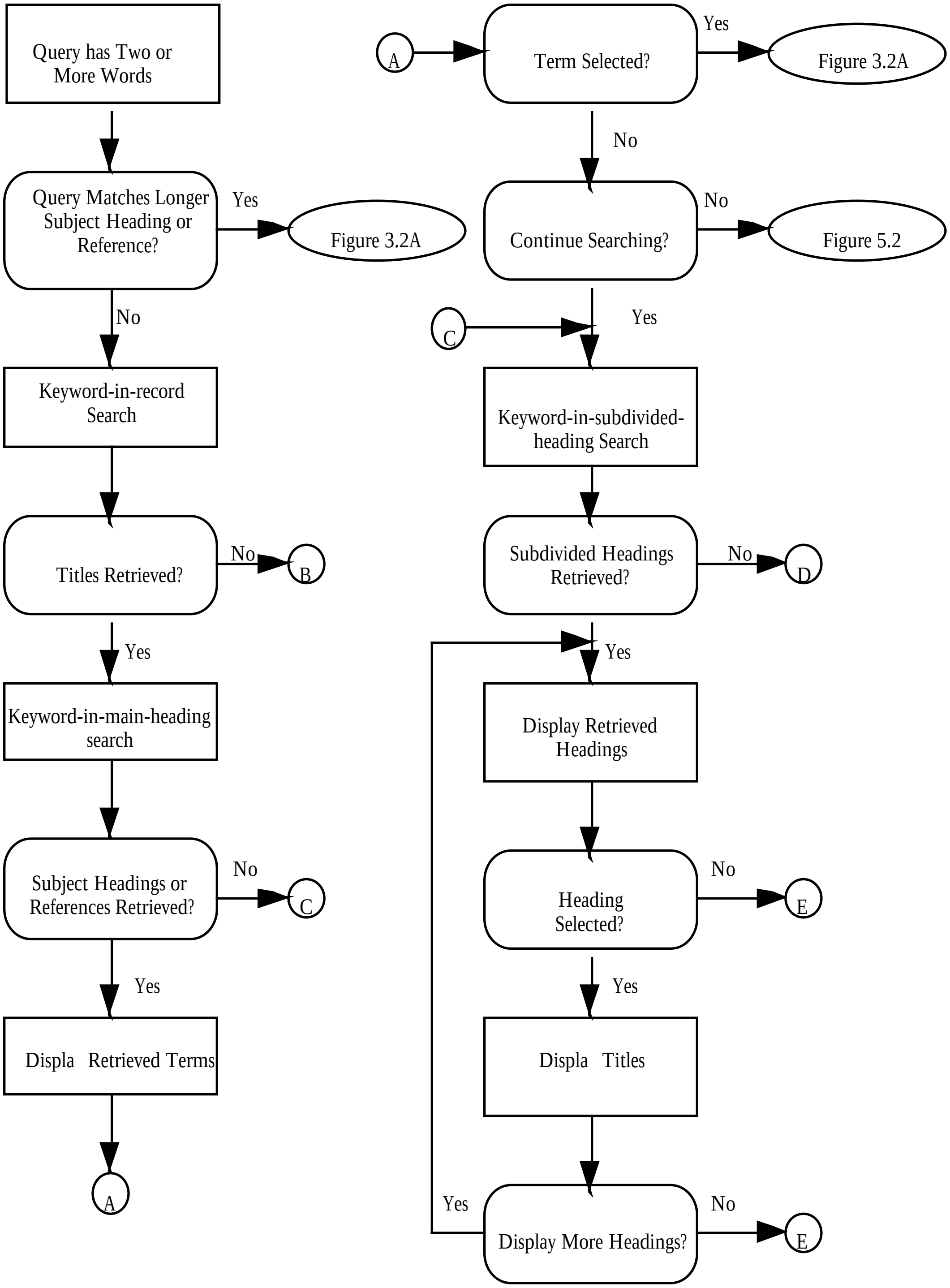

Query has Two or More Words
Query Matches Longer Subject Heading or Reference?
Yes
Figure 3.2A
No
Keyword-in-record Search
Titles Retrieved?
No
B
Yes
Keyword-in-main-heading search
Subject Headings or References Retrieved?
No
C
Yes
Displa Retrieved Terms
A
A
Term Selected?
Yes
Figure 3.2A
No
Continue Searching?
No
Figure 5.2
Yes
C
Keyword-in-subdivided-heading Search
Subdivided Headings Retrieved?
No
D
Yes
Display Retrieved Headings
Heading Selected?
No
E
Yes
Displa Titles
Yes
Display More Headings?
No
E

Figure 5.4A. Redesigned search tree for multi-word queries

If user queries failed to meet the criteria for the alphabetical approach, systems performed keyword-in-record searches for the individual words of user queries. If systems failed to produce retrievals for one or more query words, they presented the words to users for spelling correction (figure 5.4C). Systems could assist users in correcting spelling errors using spelling-correction routines similar to such routines for word processing programs in which they suggested alternate spellings for misspelled words. If users failed to correct spelling errors, systems could check queries against go/see lists to determine if queries contained listed words or phrases (figure 5.4C). They then enhanced queries with words and phrases from go/see lists and started at the initial search tree to find matches.

If the individual words in user queries produce retrievals in keyword-in-record searches, systems did not immediately show users the results (figure 5.4A). Instead, they passed queries onto other controlled vocabulary searches — the keyword-in-main-heading and keyword-in-subdivided-heading searches. If these searches produced too few or no retrievals, systems submitted queries to free-text approaches (figure 5.4B).

Free-text approaches began with the title-keyword search (figure 5.4B). If searchers wanted to find additional titles, systems could invoke relevance feedback in which they retrieved additional titles based on searches for subject headings common to titles users rated useful in title-keyword searches. If title-keyword searches failed to produce retrievals, systems could try keyword searches of subject heading fields and keyword-in-record searches. Following successful matches, systems could invoke relevance feedback using subject headings common to titles users rated useful to find additional titles. Of course, relevance feedback assumed that systems collected relevance assessments during the display of retrieved titles.

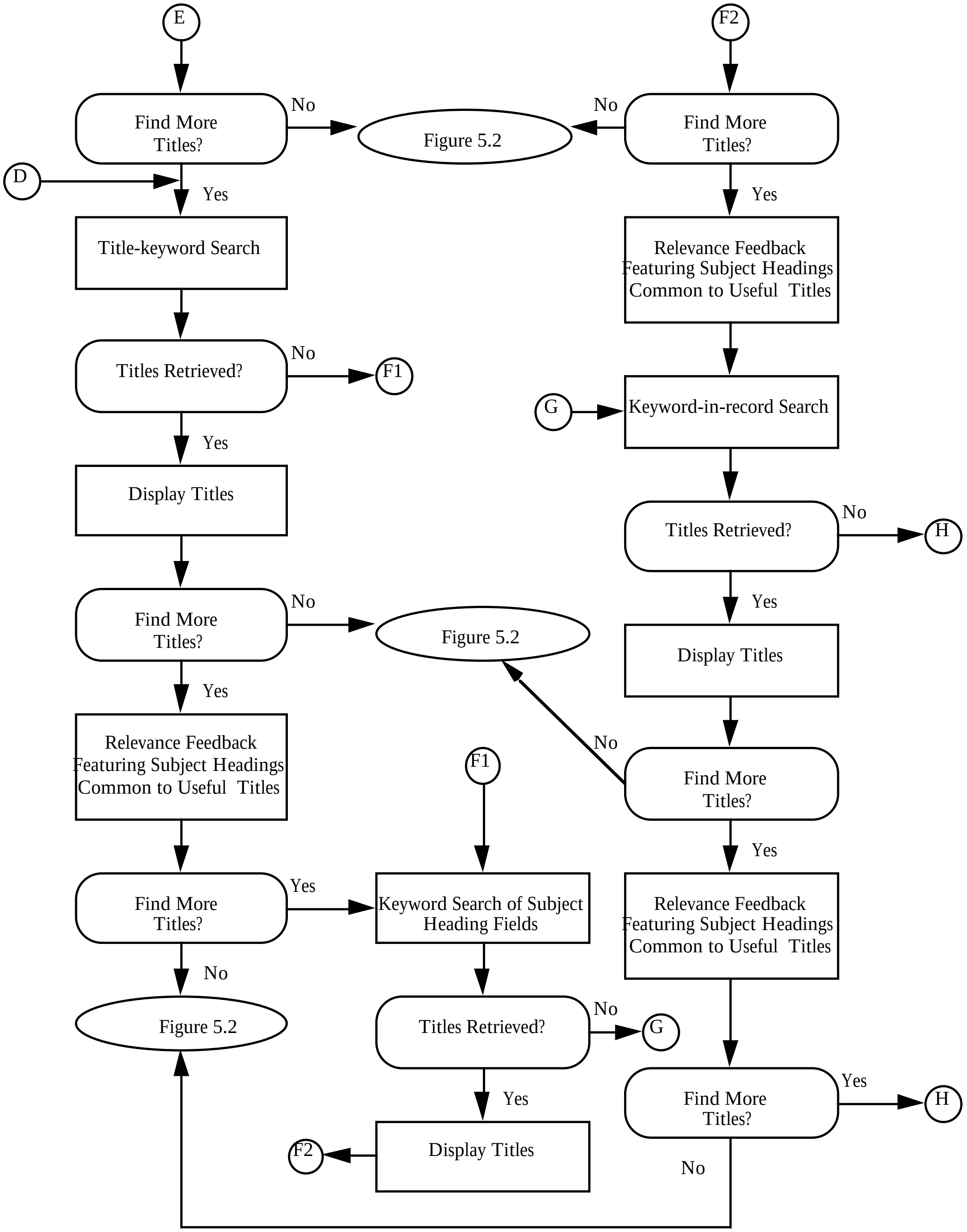


Figure 5.4B. Redesigned search tree for multi-word queries (contd.)

Following the keyword-in-record search were two new approaches (figure 5.4C). One approach involved stemming of query words. The order of indexes that systems would search for stemmed query words were exact, keyword-in-main-heading, keyword-in-subdivided-heading, title-keyword, keyword in subject heading fields, and keyword-in-record. If systems were unsuccessful making matches of user queries, they should continue searching with the best-match approach.

The best-match approach would feature stemming, weighted-term probabilistic retrieval, and output ranking. Systems would perform searches on a best match, combinatorial basis. A match on fewer than all query-word stems might retrieve titles but titles that had all the query-word stems would be displayed first.

Query-word stems would be weighted so that frequently-occurring stems would get a low "importance" weight compared to the weight assigned to rare word stems. A new weighting system should be devised for library cataloging databases because subject content is represented in so few fields of bibliographic records, i.e., title and subject heading fields. Systems should increase the weight given to frequently-occurring words in subject headings and subdivisions because matches on such words are likely to retrieve subject headings describing users' topics of interest or a facet of their topics of interest.

The alphabetical approach was added to the search tree for multi-word queries as the search type "of last resort." If all other subject searches failed to produce retrievals, systems would respond with the results of the alphabetical approach, i.e., a display of subject headings and *see* references in the alphabetical neighborhood of the user-entered term.

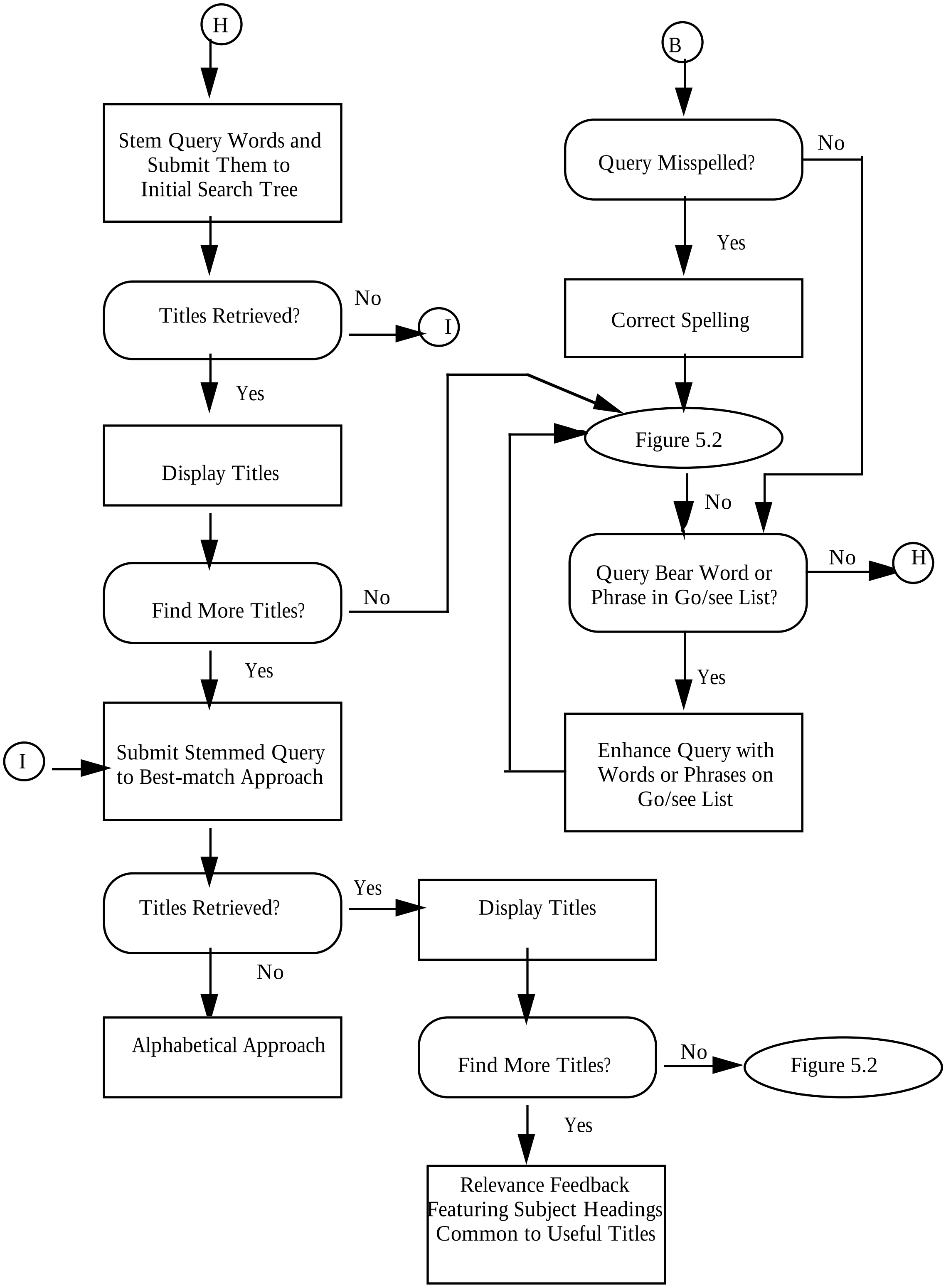

H
Stem Query Words and Submit Them to Initial Search Tree
Titles Retrieved?
No
I
Yes
Display Titles
Find More Titles?
No
Yes
I
Submit Stemmed Query to Best-match Approach
Titles Retrieved?
Yes
Display Titles
No
Alphabetical Approach
Find More Titles?
No
Figure 5.2
Yes
Relevance Feedback Featuring Subject Headings Common to Useful Titles
B
Query Misspelled?
No
Yes
Correct Spelling
Figure 5.2
No
Query Bear Word or Phrase in Go/see List?
No
H
Yes
Enhance Query with Words or Phrases on Go/see List

Figure 5.4C. Redesigned search tree for multi-word queries (contd.)

5.4Testing Enhanced Search-tree Configuration

With these changes to the original search-tree configuration in mind, we divided non-matching queries into three groups according to the number of significant words in them: (1) one-word, (2) two-word, and (3) more than two-word queries. We then determined how non-matching queries would fare in systems that were governed by the enhanced search-tree configuration.

5.4.1 One-word Queries

Only five one-word queries resulted in no matches: “kmart” (twice), “l’arche,” “transvesticism, “retibol”, and “gnatting.” The enhanced search trees would submit these queries to alphabetical searches. Of the five queries, only “transvesticism” would potentially retrieve useful titles because it would be placed in the alphabetical list of subject headings nearby “Transvestism ” and “Transvestites.” Collection failure was the reason why the query “kmart” failed to yield retrievals. Searches in business databases, i.e., Business Index and ABI/Inform, yielded 121 and 1,109 citations. Examples of retrieved citations are “A leaner, meaner Kmart,” “Retail surge leaves Kmart behind,” and “Kmart announces the details of new stock issues plans.” The queries “l’larche,” “retibol,” and “gnatting” would not be nearby potentially useful subject headings. For example, “l’arche” was nearby subject headings such as “Larch leaf-roller,” “Lard,” and “Lares.” It is doubtful users would find useful titles if they were seeking information on “L’arche,” an international organization that assists mentally-challenged adults. We failed to verify the term “retibol” in databases on social science, medicine, engineering, and business topics. The user who entered the query “gnatting” could have been looking for information on gnats or gnathology.

5.4.2 Two-word Queries

We divided the total of 36 two-word queries that failed to produce retrievals in an online bibliographic system governed by the original search-tree configuration into three main groups: (1) queries for which the enhanced search trees produced retrievals in library-catalog databases, (2) queries for which the enhanced search trees produced retrievals in subject-specific databases, and (3) queries which still resulted in no retrievals or no seemingly useful titles. (Two-word queries included three-word queries in which one of the three words was a stopword.) Table 5.2 summarizes examples of the former.

Table 5.2. Matches of Two-word Queries

User Query	Explanation

smoking woman	The inclusion of irregular plurals in a go/see list would produce two dozen potentially useful retrievals for the manipulated query “smoking women.” Examples are *Women and substance abuse, Women and tobacco,* and *Effects of smoking on the fetus, neonate, and child*. These titles also contain useful subjects headings e.g., “Women — Tobacco use” and “Pregnant women — Tobacco use.”
directories retailers	Truncation results in keyword-in-heading matches of the subdivided heading “Retail trade — Directories.” Examples of potentially useful titles are *World-wide franchise directory* and *Michigan distributors directory*.
federalisms and jefersonianism	Truncation and spelling correction results in keyword matches. Examples of potentially useful titles are *Federalists in dissent: imagery and ideology in Jeffersonian America* and *The Federalist Party in the era of Jeffersonian democracy*.
rome oly	Truncation results in keyword-in-heading matches. Examples of potentially useful titles are *The games of the XVII Olympiad, Rome* and *Olympiad 1960*.

Examples of queries that produce retrievals in subject- or form-specific databases are listed in Table 5.3.

Table 5.3. Matches of Two-word Queries in Subject- or Form-specific Databases

User Query	Match type/Database
golgi body	keyword-in-record/Medline
technique in neutralizations	keyword-in-record/Medline
stuttering psychogenic	keyword-in-record, truncation/Medline
bulimia test	keyword-in-record/Medline
glycolysis and respiration	keyword-in-record/Medline
lighting concentration	keyword-in-record, truncation/Compendex
transvection effects	keyword-in-record, truncation/Medline
engine keys	keyword-in-record/Compendex
flamability standards (sic)	keyword-in-record, spelling correction/Compendex
keystone corporation	keyword-in-record/ABI/Inform
racism in humor	keyword-in-record/National Newspaper Index

Despite truncation, spelling correction, and searches in subject-specific databases, twenty two-word queries still failed to produce retrievals. If systems responded with alphabetical searches, surprisingly, most of the searches would list potentially useful subject headings in alphabetical lists; however, users would have to be patient and persevering to browse backward and forward in alphabetical lists of subject headings to find suitable matches. Table 5.4 lists these queries of potentially useful subject headings.

Table 5.4. Alphabetical Matches of Two-word Queries

Query	Potentially useful subject headings
biodiversity and virginia	Biodiversity, *see* Biological diversity
black playwrites	Black Americans, *see* subject headings beginning with the words Afro-American
boxing tragedies	Boxing, Boxing — Accidents, *see* Boxing — Accidents and injuries
cardiovascular videocassettes	Cardiovascular agents; Cardiovascular emergencies; Cardiovascular system; etc.
cremastogaster pilosa	Cremastogaster
critique of ironwood	
film rationing	Film, Motion picture, *see* Motion picture films; Film industry (Motion pictures), *see* Motion picture industry
functionalism and media	Functionalism (Psychology); Functionalism (Social sciences); etc.
geisha dancers	Geishas
good furniture	
islamic revitalization	Islamic countries; Islamic fundamentalism; etc.
neander valley	Neanderthals
penis expansion	Penis
psychosis icu	Psychoses
race g	Race; Race discrimination; Race prejudice, *see* Racism; etc.
rayonnant architecture (twice)	
speech critiques	Speech; Speeches, addresses, etc.; Speech criticism, *see* Rhetorical criticism
toshiba affair	Toshiba Strike, 1949
truckbed liners	Trucking; Trucks; etc.

In Table 5.4, only three queries are not accompanied by a potentially useful subject heading, i.e., “critique of ironwood,” “good furniture,” and “rayonnant architecture.”

5.4.3 Queries Exceeding Two Words

Remaining were 49 queries. These queries exceeded two words and failed to produce retrievals in an online bibliographic system governed by the original search-tree configuration. They could be divided into three main groups: (1) queries for which the enhanced search trees produced retrievals in library-catalog databases, (2) queries for which the enhanced search trees produce retrievals in subject- or form-specific databases, and (3) queries which still result in no retrievals or no seemingly useful titles. Table 5.5 summarizes examples of the former. It includes examples of queries that produce matches through the best-match approach. Terms ending in a plus sign (+) designate where truncation was applied.

Table 5.5. Matches of Queries Exceeding Two Words

User Query	**Explanation**
architecture in the 16th to18th centuries (sic)	Best-match approach on the truncated words "architecture 16th centur+" retrieves titles such as *Renaissance architecture*, *The royal palaces of Tudor England*, *Half-timbered houses and carved oak furniture of the 16th and 17th centuries*, *Elizabeth's England*, and *The cottages of England*.
collegiate sports recruiting	Keyword matches on the truncated words "colleg+ sport+ recruit+" retrieve titles such as *The recruiting game: toward a new system of intercollegiate sports* and *The outside shot*.
legal cases involving life support systems in humans	Best-match approach on the words "legal life support" retrieves titles such as *Abatement treatment with critically ill patients: ethical and legal limits to the medical prolongation of life*, *The right to die*, and *No heroic measures: moral, ethical, and legal issues in the neurosciences*.
love and marriage in islam	Best-match approach on the words "love islam" or "marriage islam" yields dozens of titles.
social problems and drugs and unemployment	Best-match approach on the words "social problems drugs" or "social problems unemployment" yields dozens of titles.
ra reading aloud to children	Best-match approach on the words "reading aloud children" retrieves the title *Stories in the classroom: storytelling, reading aloud, and roleplaying with children*.
iraqi political systems	Keyword-in-heading matches on the truncated words "iraq+ politic+ system+" retrieve several titles bearing the subject heading "Iraq — Politics and government."
french occupation in chad	The inclusion of proper adjectives in a go/see list would produce potentially useful retrievals for the manipulated query "france occupation in chad" in a keyword-in-record search.

severly physically handicapped students	Spelling correction and best-match approach on the truncated words "sever+ physical+ handicap+" retrieve titles such as *Adjustments to severe physical disability*, *Severe disabilities*, and *Handbook of severe disability*.
jews in post war era	Best-match approach on the words "jew+ post war" retrieves titles such as *Living after the Holocaust: reflections by the post-war generation in America*, *The future of the Jews*, and *Jews in the post-war world*.

Examples of queries that produced retrievals in subject- or form-specific databases are listed in Table 5.6.

Table 5.6. Matches of Queries Exceeding Two Words in Subject- or Form-specific Databases

User Query	Match type/Database
lsi logic corporation	keyword-in-record, truncation/ABI/Inform
capital punishment and terrorists	keyword-in-record, truncation/National Newspaper Index
myeloid cell lines	title-keyword/Medline
congenital dyserythropoietic anemia	title-keyword/Medline
sexual abuse in poverty families	keyword-in-record, truncation/Psychological Abstracts

Although the best-match approach was successful in retrieving titles, it did not always address the topics expressed in queries. Examples are given in Table 5.7.

Table 5.7. Retrievals from Best-match Searches

User Query	Match type/Database
suicide and mass communication	Titles retrieved in best-match searches with the words "mass communication" are too general and with the words "mass suicide" are false drops.
women in the work force in the 1900–1950	Best-match approach on the words "women work force" produces titles but none that specifically address this time period.
kodak disc camera	Best-match approach on the truncated words "kodak camera+" yields a dozen titles including *The story of Kodak* and *History of Kodak cameras*. User must review contents and index to find references to the "kodak disc camera."

coins of the united states in 1943	The inclusion of proper adjectives in a go/see list would produce two dozen potentially useful retrievals for the manipulated query “coins american” through the best-match approach. Examples are *A history of United States coinage* and *Coin world: comprehensive encyclopedia of United States coins.* Users must review contents and index to find references to 1943.
labor unions and depression and michigan	Best-match approach yields retrievals for three of four substantive words, e.g., “labor union+ michigan” or “labor union+ depression.” Failure to make matches of all four substantive words results in titles that are not specific to the user’s interests.
holistic health creative visualization	Best-match approach yields retrievals for “holistic health” or “creative visualization.” User wants both topics addressed in a single work.
gift wrapping machines	Best-match approach yields retrievals for “gift wrapping” or “machines.” User wants both topics addressed in a single work.

The best-match approach would produce retrievals for all queries exceeding two words. The problem is some retrievals will be based on matches on one or two truncated words in lengthier queries. Consequently, retrievals will miss important facets of the topics expressed in user queries. If alphabetical searches are enlisted as the approach “of last resort,” the same thing happens, that is, nearby matches of subject headings in alphabetical lists express one facet of users’ multi-faceted queries. Examples of queries for which no amount of manipulation produces useful results are:

- law predictive theory
- labor market cartels
- mathematic theory correctness
- diagnosis clinical popular
- graphic intyerchane files (sic)
- pressure temperature windchield factor (sic)

5.5Chapter Summary

Of the 1,919 queries for subjects generally, 90 (4.7%) failed to meet the criteria for exact, alphabetical, keyword-in-heading, and keyword matches (section 5.2). The majority (78.9%) of non-matches were queries for topical subjects (Table 5.1). Non-matches figured into lengthy searches that averaged four access points per search. They averaged over three words per query and failed to retrieve titles through the existing search-tree configuration. Chapter 5 takes an in-depth look at these queries, suggests ways in which this configuration could be enhanced to handle them, and tests the enhanced configuration using the non-matching queries in this study.

We used an enhanced search-tree configuration from a related study to determine its effectiveness responding with useful retrievals to queries for which the original search-tree configuration was unable to produce retrievals (Drabenstott and Weller 1994, chapter 13).

The initial search tree remained basically unchanged from the original initial search tree (figure 5.2). The only change was to return users whose searches failed to produce retrievals to the question about personal names. Redesigned search trees for one-word and multi-word queries contained several changes from the original search trees: (1) checking go/see lists, (2) redirecting misspelled queries to the question on personal names, (3) invoking the keyword-in-record search (added to one-word tree), (4) invoking relevance feedback following keyword searches (i.e., title-keyword, keyword in subject heading fields, and keyword-in-record searches), (5) invoking stemming (multi-word search tree only), (6) invoking the best-match approach (multi-word search tree only), and (7) invoking the alphabetical approach as the search type “of last resort” (figures 5.3A–5.3B and 5.4A–5.4C).

With these changes to the original search-tree configuration in mind, we divided non-matching queries into three groups according to the number of significant words in them: (1) one-word, (2) two-word, and (3) more than two-word queries. We then determined how non-matching queries would fare in systems that were governed by the enhanced search-tree configuration (section 5.4).

One-word and several two-word non-matching queries showed promise in terms of retrieving useful information through alphabetical searches. Truncation, spelling correction, and the availability of a go/see list with irregular plurals combined to produce retrievals for several two-word queries. Truncation, spelling correction, the availability of a go/see list with proper adjectives, and best-match approach combined to produce satisfactory retrievals for several queries that exceeded two words. Searches of subject- or form-specific databases (e.g., Medline, Psychological Abstracts, National Newspaper Index) for several non-matching queries composed of two or more significant words was also a successful strategy for producing retrievals. No amount of manipulation produced useful results for a small number of queries.

5.6References

Drabenstott, Karen M., and Marjorie S. Weller. 1994. *Testing a new design for subject searching to online catalogs.* Ann Arbor, MI: School of Information and Library Studies, University of Michigan.

6 Queries for Personal Names

6.1 Introduction

One of the key findings from the original empirical study of end-user queries was the definition of different search trees for queries for subjects generally and for queries for personal names (Drabenstott and Vizine-Goetz 1994, 230–3). The search trees differed because they took into account the content and structure of assigned subject headings for personal names.

Personal name headings come from the assigned subject headings in libraries' bibliographic records. Catalogers establish these headings by following AACR2 rules and guidelines and previous LC practice. They refer to the Library of Congress Name Authority File (LCNAF) to verify personal names. References are generated from LCNAF records for names used in libraries' bibliographic records. Catalogers also refer to the *Subject cataloging manual: subject headings* (SCM:SH) to add subdivisions to personal name headings. The results are single assigned subject headings bearing elements for personal names, topical subjects, and, possibly, geographic names. The order of elements in such headings are: (1) surname, (2) given name or initial, (3) possibly, a middle name or initial, (4) possibly, date or date range, (5) possibly, topical subdivisions, and (6) possibly, geographic subdivisions. Subdivisions appended to personal name headings are usually for topical subjects; however, several subdivisions under names of persons authorize use of geographic subdivisions (Library of Congress 1990, H1110).

The original empirical study demonstrated that users rarely expressed all six elements of personal name headings in their queries. Furthermore, when two or more elements were included, users rarely entered elements in the sequence prescribed by personal name headings.

When searching for user queries bearing personal name elements, systems need information about the particular elements included in queries to distinguish personal name elements from topical subject elements. Systems could then accommodate the rigid structure of personal name headings by searching for as many elements as needed to provide useful retrievals.

None of the four systems from which we obtained transaction log data specifically prompted end users to enter topical and personal-name elements of their queries. SULIRS, LS/2000, and MIRLYN handled subject queries bearing personal-name elements the same as queries for subjects generally. ORION required users to enter subject queries bearing personal-name elements with a command that was different from the command for subjects generally; however, this system did not prompt users for personal-name or topical elements of their queries.

To perform the analysis described in this chapter, we reviewed user queries to determine whether they contained personal name elements. If they contained such elements, we identified first name, middle name, last name, and topical elements, and submitted them to search trees for personal-name queries (Table 3.1). In an operational system governed by search trees, systems would ask users whether their queries involved a personal name and prompt them for the various elements.

6.2 Combinations of Topical and Personal-Name Elements

6.2.1 Introduction

Search trees for personal-name queries first determined whether queries contained topical elements in addition to one or more personal-name element(s). If they had topical elements, systems submitted queries to keyword-in-subdivided-heading searches to find titles bearing both name and topical elements. They even omitted middle and first name elements to effect keyword-in-heading matches. Table 6.1 is a fact sheet for user queries for personal names that contained topical and personal-name elements. These queries were candidates for keyword-in-heading and keyword-in-record searches.

Table 6.1. Fact Sheet for Personal-name Queries Bearing Topical and Name Elements

Facts	Total	SULIRS	ORION	LS/2000	MIRLYN
Searches					
No. of searches	32	13	3	5	11
No. of access points	105	48	4	14	39
Avg. no. of access points per search	3.3	3.7	1.3	2.8	3.5
Maximum no. of access pts. in a search	13	13	2	9	13
Initial access points					
Avg. no. of words per access point	3.4	2.9	3.0	3.8	4.0
Maximum no. of words in an access pt.	10	5	4	5	10
Avg. no. of retrievals per access point	N/A	8.0*	0.7†	N/A	N/A
Percentage of access points with zero retrievals	N/A	61.5	66.7	N/A	36.4

Percentage of access points with retrievals > 999	N/A	0.0*	0.0†	N/A	N/A

*Number of bibliographic records retrieved. This was an estimate based on substituting 3,000 retrievals for the 99,999 retrievals that was written to logs when number of retrievals exceeded about 1,000.
†Number of assigned subject headings retrieved.

Of the total of 251 queries for personal names, only 32 (12.8%) contained topical elements. Thus, queries bearing personal-name and topical elements were not very common in the logs we reviewed. Except for ORION (which contributed only three queries of this type), users entered two or more queries following their initial query. Some searches were quite long. For example, a MIRLYN user entered twelve more queries following his initial query “skinner and sibling\s” (sic). Subsequent queries demonstrated the user’s interest in siblings, e.g., “skinner and siblings,” “gesell and siblings,” “siblings and the oedipal complex,” and “theoretical approach to sibling differences.”

Queries bearing personal-name and topical elements were quite long; they averaged 3.4 words per access point. Of course, this type of query required at least two access points: a name element and a topical element. In SULIRS, postings were quite low. They were even lower in ORION because this system featured a separate command for the entry of queries for personal names (section 2.2.3).

6.2.2 Keyword and Alphabetical Matches

Search trees for personal names first tested queries to determine if they were candidates for keyword-in-heading or keyword-in-record searches. Thus, our analysis of user queries for personal names continued with this test. We separated personal-name queries bearing topical and personal-name elements from those bearing only personal-name elements. Table 6.2 shows the result.

Table 6.2. Types of Keyword Matches

Type of match	Personal-name searches	Subjects generally
Keyword-in-record	5	0
Keyword-in-subdivided-heading	2	1
Alphabetical	12	1
Exact	N/A	4
Title	N/A	5
None	0	2
Total	19	13

A total of 32 queries contained both topical and personal-name elements. What was particularly surprising about this analysis was that almost as many queries bearing personal names should have been submitted to search trees for subjects generally as to search trees for personal names. Examples of these queries are:

- keyenesian economics (sic)
- the hapsburg monarchy
- notions of buddha
- myers-briggs type indicator
- hannibal and the battle of carthage
- ras tafari

These and the several other queries that involved personal names presented a problem. They contained personal-name and topic elements but they would have been satisfied by subject searches in the search trees for subjects generally. End users would not know (and should not *have* to know) that these queries would be satisfied by the search trees for subjects generally.

An experimental online catalog governed by search trees asked searchers the question "Does your query involve a personal name?" to help it differentiate the two types of queries. Users occasionally answered this question positively when they should have answered it negatively (Drabenstott and Weller 1994). Additional research is needed to help users and systems differentiate queries for subjects generally from queries for personal names.

A handful of personal-name queries were keyword matches. Examples are:

- clinton and poverty
- freud and aggression
- senator lloyd bentsen
- religion tolstoy
- chaucer criticism
- delacroix and colr (sic)

The remainder of personal-name queries were alphabetical matches. That is, the combination of personal-name and topical elements failed to produce retrievals so systems governed by search trees would omit the latter element and submit the remaining personal-name element(s) to alphabetical searches. Examples are the personal-name elements in the following queries:

- skinner and sibling\s (sic)
- foreign policy of roosevelt theodore
- the life of william faulkner

- descartes future prediction
- paintings of pollock
- clarence darrow's relegious views (sic)
- greek mythology's influence on shakespeare

Since alphabetical searches only considered name element(s), users would not necessarily retrieve titles on the specific topics cited in their queries. For example, searches on the "foreign policy of roosevelt theodore" retrieved titles on many different topics; users would have to persevere to find the several titles that mentioned his foreign policy, e.g., *Roosevelt and the Caribbean, Velvet on iron: the diplomacy of Theodore Roosevelt, Roosevelt and the Russo-Japanese War: a critical study of American policy in Eastern Asia in 1902–5,* and *Theodore Roosevelt: confident imperialist.*

6.3Alphabetical Matches

6.3.1 Introduction

Remaining personal-name queries contained only personal-name element(s). Table 6.3 is a fact sheet for these user queries. In an online system governed by search trees, these queries would be submitted to alphabetical searches (Table 3.1).

Table 6.3 Fact Sheet for Personal-name Queries Bearing Name Elements Only

Facts	Total	SULIRS	ORION	LS/2000	MIRLYN
Searches					
No. of searches	219	75	30	66	48
No. of access points	500	143	48	200	109
Avg. no. of access points per search	2.3	1.9	1.6	3.0	2.3
Maximum no. of access pts. in a search	16	12	6	16	14
Initial access points					
Avg. no. of words per access point	1.8	1.7	1.6	1.7	2.1
Maximum no. of words in an access pt.	6	5	4	3	6
Avg. no. of retrievals per access point	N/A	174.5*	1.0†	N/A	N/A
Percentage of access points with zero retrievals	N/A	20.0	86.7	N/A	29.2
Percentage of access points with retrievals > 999	N/A	1.3*	0.0†	N/A	N/A

*Number of bibliographic records retrieved. This was an estimate based on substituting 3,000 retrievals for the 99,999 retrievals that was written to logs when number of retrievals exceeded about 1,000.

†Number of assigned subject headings retrieved.

Of the total of 251 queries for personal names, 219 (87.3%) contained personal-name elements only. Searches ranged from one to three access points. Some searches were quite long. For example, a MIRLYN user entered thirteen more queries following his initial query “hart, gary.” Subsequent queries demonstrated the user’s interest in this person’s bid for the presidency, e.g., “extra marital affairs,” “sex,” “scandals,” “sex and presidential candidates,” “mass media,” “media coverage of elections,” “presidential scandals,” and “sex in politics.”

Personal-name access points bearing personal-name elements were only about two words long. A few queries were five or six words long. Examples of these long personal-name queries are:

- strunk w oliver william oliver 1901
- mubarak muhammad husni 1928
- prosser walter lee
- van der velde
- dr martin luther king jr
- frank lloyd wright

6.3.2 Form of Name Entered

Subject searching for personal names was handled in three different ways in the four online catalogs from which transaction log data were obtained for this study. MIRLYN’s subject heading (s=) search and LS/2000 featured two-step approaches which required users to enter inverted forms of names, e.g., “shakespeare, william,” or “twain mark,” or enter the surnames, e.g., “clinton” or “gauguin.” MIRLYN and LS/2000 responded with an alphabetical list of assigned subject headings in alphabetical proximity to the term entered. Users who selected listed subject headings would retrieve a display of bibliographic records assigned the selected heading. ORION’s keyword-in-subdivided-heading search for personal names was also a two-step process. The system responded to user-entered queries for names with a list of assigned subject headings bearing the words in user queries. Users who selected listed subject headings would retrieve a display of bibliographic records assigned the selected heading.

MIRLYN’s keyword (k=) search and SULIRS’ keyword (lc; or sb;) search were keyword-in-record searches; consequently, the order of elements in personal name queries did not matter. Keyword-in-record searches were one-step approaches in which systems retrieved bibliographic records bearing the entered terms. When users entered surnames using terms that were also given names, e.g., “grant,” “kelly,” “thomas,” the number of retrieved records could be quite high because SULIRS and MIRLYN retrieved records bearing these words irrespective of their use as given names or surnames.

Figure 6.1 summarizes the forms of names entered into the three online catalogs featuring keyword searching, viz. SULIRS, ORION, and MIRLYN, and the two online catalogs featuring alphabetical searching, viz. LS/2000 and MIRLYN.

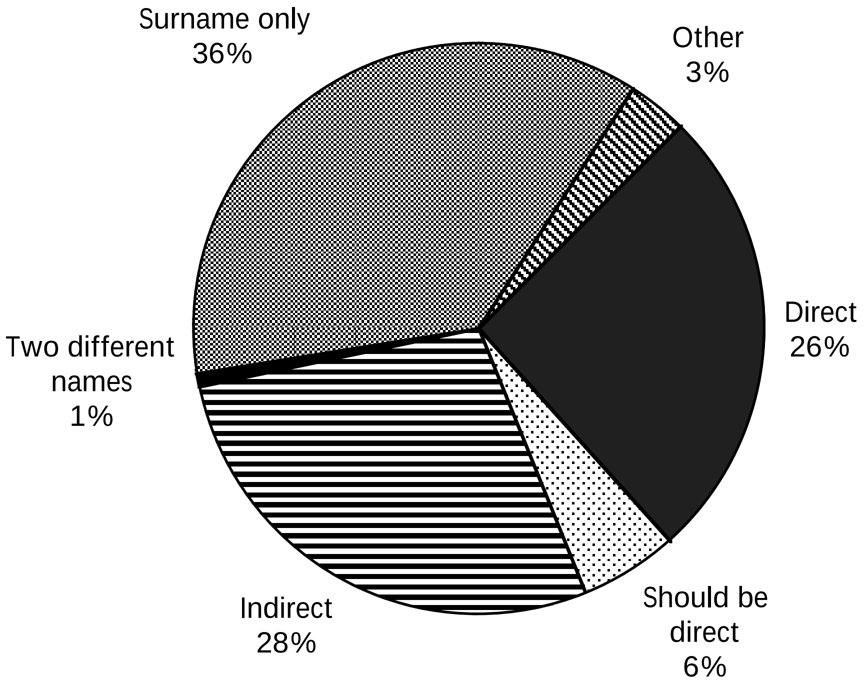


Figure 6.1. Forms of personal-name queries

A little over a third of access points consisted of surnames only. Examples are:

- lawrence
- hitchcock
- fukasa
- shakespeare
- von sternberg

Indirect forms of names were quite common, accounting for between a quarter and a third queries of personal-name queries. Examples are:

- hart, gary

- woolf virginia
- corneille, pierre
- abrahams r d
- wittgenstein, ludwig

Direct forms of names were also quite common, accounting for about a quarter of personal-name queries. Examples are:

- gertrude stein
- william james
- shinon peres
- jackson pollack (sic)
- reginald wright kauffman

A few names included dates; all but one name were indirect forms entered into MIRLYN. Examples are:

- strunk w oliver william oliver 1901
- mubarak muhammad husni 1928
- marshall george c george cutlett 1880–1959
- swift, jonathan, 1667–1745
- bach alexander 1813 1893

Queries with dates were so similar to the content, order, and format of assigned subject headings that it was very likely that users entered tracings that were displayed on bibliographic records. For example, the query “strunk w oliver william oliver 1901” was probably the qualified name in the assigned subject heading “Strunk, W. Oliver (William Oliver), 1901–” of a bibliographic record that was displayed on the screen while the user typed in this personal-name query. MIRLYN users were probably more predisposed than users of the other three systems to include dates in their personal-name queries because this system prompted users to enter tracings listed on bibliographic records. The only other query bearing dates was entered by an ORION user (last-listed query with dates above).

About 6% of personal names entered by online catalog users were in and should have been in direct form. Examples are “augustine,” “aristotle,” “black elk,” “ovid,” and “el greco.”

Two queries consisted of two names. These were connected with the Boolean “and” operator and came from LS/2000 which did not feature an explicit “and” operator: “nietzche and kierkegard” and “sacco and vanzetti.” A discussion of the names placed in the “other ” category is interesting. One name consisted only of initials, i.e., “fdr.” LCNAF provided a *see* reference under these initials to direct users to the authorized form of name. The given name "vincent" was entered in a search in which the user was probably interested in the artist "Vincent van Gogh." Searches in the LCNAF and several online catalogs failed to verify "n scribner richard," "yassin adj ramadan," and "faisal ibn 'abd alaziz." A *see* reference would have helped the user entering "le roi soliel" to find information on the French King Louis XIV.

6.3.3 Alphabetical Approach for Personal-name Queries

In bibliographic systems governed by search trees, the alphabetical search was the recommended approach for personal name queries without topic elements. This search placed users in the alphabetical index of personal names where their entered names were or would be listed in the alphabet. Systems always provided a response to subject searches for personal names through the alphabetical approach. It is important that users entered the correct element for last names because the system's placement at a particular point in the alphabetical index nearby the desired personal name depended upon this element.

To perform the analysis of personal-name queries, we reviewed user queries to determine whether they contained personal name elements (section 6.2.1). If they contained such elements, we identified first name, middle name, last name, and topical elements, and submitted them to the search tree for personal-name queries (Table 3.1). Remaining queries were submitted to the alphabetical approach. We used our judgment in reordering the name elements of direct forms of personal-name queries to reflect the order of name elements in assigned subject headings for personal names. In an operational system governed by search trees, systems would ask users whether their queries involved a personal name, prompt them for the various elements, and use the surname and given name elements to place users in the appropriate location in the alphabetical name index. Table 6.4 shows the result of our analysis of end-user queries.

Table 6.4. Types of Alphabetical Matches

Type of match	Personal-name searches	Subject searches generally
Alphabetical	211	1
Exact	N/A	4
Keyword-in-main-heading	N/A	3
Total	211	8

Systems governed by search trees would submit a total of 211 queries to the alphabetical approach for personal-name queries. They would also submit eight searches that should have been submitted to various searches controlled by the search tree for subjects generally to the alphabetical approach for personal-name queries because these eight queries involved personal names. Users would not have been satisfied with the result because the desired name would not have been listed on the initial, nearby, or distant screens of alphabetically-arranged subject headings for personal names. Examples of these queries and matching topical subject headings are:

User Query	Matching heading	Subject search
odysseus	Odysseus (Greek mythology)	Exact
aphrodite	Aphrodite (Greek deity)	Exact
philoctetess (sic)	Philoctetes (Legendary character)	Exact
oedipus	Oedipus (Greek mythology)	Exact
nmarlowe philip (sic)	Marlowe, Philip (Fictitious character)	Alphabetical

These names were connected with mythological, fictitious, or legendary characters. Systems that prompt end users for elements of personal-name queries or handle personal-name queries separately from queries for subjects generally must experiment with the wording of questions that ask users to distinguish between types of queries. An even better solution would be for systems to process subject headings bearing qualifiers such as “(Legendary character),” “(Fictitious character),” “(Greek deity),” and “(Roman deity)” into indexes for both personal names and subjects generally.

6.3.4 Match Satisfaction

Figure 6.2 shows satisfaction categories for alphabetical matches. Percentages do not include the eight searches involving names that occurred in topical subject headings. These searches should have been submitted to the search trees for subjects generally.

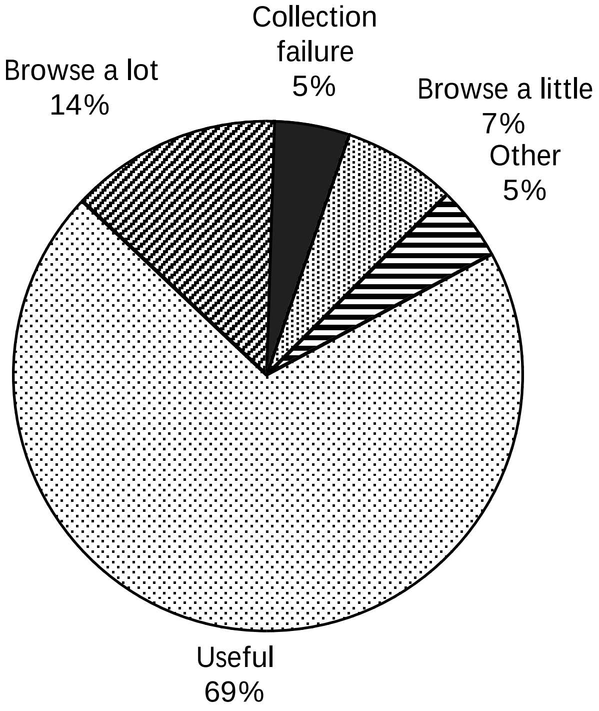


Figure 6.2. Satisfaction categories for alphabetical matches

Over two-thirds of alphabetical matches would lead users to personal-name subject headings on their topics of interest. Examples of satisfactory alphabetical matches of subject headings and references are:

User query	Matching subject heading
corneille, pierre	Corneille, Pierre, 1606–1684
laban	Laban, Rudolf von, 1879–1958
twain, mark	Twain, Mark, 1835–1910
augustine	Augustine, Saint, Bishop of Hippo
swift, jonathan, 1667–1745	Swift, Jonathan, 1667–1745
de genlis	De Genlis, Stéphanie Félicité, comtesse, 1746–1830 (reference)

Users must browse for about 20% of their desired subject headings. The “browse a little” category meant users would find their desired subject headings on the screen immediately preceding or following an initial screen of a dozen or fewer subject headings for personal names. The “browse a lot” category meant users would have to browse more than one screen. For example, the user who entered the query “grant” would have had to browse through many names before arriving at the “U’s” where the desired subject heading “Grant, Ulysses S.

(Ulysses Simpson), 1822–1885" resided. Other examples of queries that would require a little or a lot of browsing to reach the desired personal-name subject heading are:

User query	Matching subject heading
lawrence	Lawrence, D. H. (David Herbert), 1855–1930
neera	Neera, 1846–1918
torres	Many names beginning with "Torres"
clinton	Clinton, Bill, 1946–
co prbusier (sic)	Corbusier, 1887–1965 (reference)
kelly	Many names beginning with "Kelly"
marlowe, jukli{ (sic)	Marlowe, Christopher, 1564–1593
jackson pollack (sic)	Pollock, Jackson, 1912–1956
butler	Many names beginning with "Butler"
farakan louis (sic)	Farrakhan, Louis

Several personal-name queries were common surnames, e.g., "lawrence," "butler," "kelly." A few personal-name queries that consisted of first- and last-name elements and required much browsing had spelling errors, e.g., "co prbusier, " "jackson pollack," and "farakan louis." Such errors typically resulted in the display of an alphabetical list of personal-name subject headings that did not place users close to the desired heading; consequently, they might have had to do a considerable amount of browsing to find the headings they wanted.

When we were unable to verify names in searches of several online catalogs, we selected the "collection failure" satisfaction category. Examples of these queries are:

- klagsburn
- steinway henry
- marie teppp
- n richard scribner
- abrahams r d

6.4 Chapter Summary

Search trees for queries for personal names differed from search trees for queries for subjects generally because the former took into account the content and structure of assigned subject headings for personal names. When searching for user queries bearing personal name elements, systems need information about the particular elements included in queries. They would use this information to accommodate to the rigid structure of personal name headings by searching for as many elements as needed to provide useful retrievals. To perform the analysis of personal-name queries described in chapter 6, we reviewed user queries to determine whether they contained personal name elements. If they contained such elements, we identified first name, middle name, last name, and topical elements, and submitted them to search trees for personal-name queries (Table 3.1). In an operational system governed by search trees, systems would ask users whether their queries involved a personal name and prompt them for the various elements.

Of the total of 251 queries for personal names, only 32 (12.8%) contained topical elements (section 6.2). Queries bearing elements for topics and personal names amounted to only 1.7% of the queries in this study (Table 6.1). Queries bearing personal-name and topical elements were quite long averaging 3.4 words per access point. This type of query required at least two access points: a name element and a topical element. Of the total of 32 queries that contained both topical and personal-name elements, almost as many queries bearing personal names should have been submitted to search trees for subjects generally as to search trees for personal names (Table 6.2). End users would not know (and should not *have* to know) that these queries would be satisfied by the search trees for subjects generally. Additional research is needed to help users and systems differentiate queries for subjects generally from queries for personal names.

A handful of the queries bearing topics and personal names were keyword matches. Queries that were not keyword matches were submitted to the alphabetical approach minus the topic element(s). Since alphabetical searches only considered name element(s), most alphabetical searches would require users to persevere to find promising titles on the specific topics that interested them.

A total of 219 queries contained only personal-name element(s) (Table 6.3). In an online catalog governed by search trees, these queries would be submitted to alphabetical searches (section 6.3). Although two catalogs had strict guidelines about the order of entry elements in personal-name queries, users entered personal-name elements in every order imaginable, e.g., direct, indirect, acronyms, surnames only, given names only (figure 6.1). Eight queries should have been submitted to the search trees for subjects generally because they named mythological, fictitious, or legendary characters. Systems should process subject headings bearing qualifiers such as “(Legendary character),” “(Fictitious character),” “(Greek deity),” and “(Roman deity)” into indexes for both personal names and subjects generally so that end users would retrieve the information they wanted regardless of their responses to the system’s initial question about whether their queries involved a person’s name.

Over two-thirds of alphabetical matches would lead users to personal-name subject headings on their topics of interest (figure 6.2). Users must browse for about 20% of their desired subject headings.

6.5References

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

Drabenstott, Karen M., and Marjorie S. Weller. 1994. *Testing a new design for subject searching to online catalogs*. Ann Arbor, MI: School of Information and Library Studies, University of Michigan.

Library of Congress. 1990. *Subject cataloging manual: subject headings*. Washington, D.C.: Library of Congress.

7 Classification for Subject Searching

7.1 Introduction

A large share (43.4%) of end-user queries were exact matches of controlled vocabulary terms. Unfortunately, exact matches produced too many retrievals. Exact matches of topical subjects and geographic names were especially plagued by too many retrievals. Too many retrievals were reported for 28.6% of exact matches of the former and for 60.7% of exact matches of the latter (Table 4.8). The exact approach was recommended as the system response to exact matches. This approach anticipated the user's selection of the exact match from an alphabetical list, and thus began with a report of the results of such exact matches. The report included a summary of subdivided forms of the matched subject heading, and, if available, options for browsing related terms and other information about the matched heading.

The exact approach could be implemented in existing online catalogs; however, without further developmental work, it would be limited in usefulness because only three broad categories could be constructed from the three different types of subject subdivisions coded in subject heading fields of bibliographic records. Drabenstott and Vizine-Goetz (1994, 254–60) describe the editorial and developmental work efforts needed to fully implement the exact approach in online bibliographic systems.

Alternatives to additional subfield codes for managing the many subdivided forms of assigned subject headings can be drawn from the suggestions of researchers who have been studying the problem of large retrievals (Larson 1989; Prabha 1990; Lynch 1990; Wiberley, Daugherty, and Danowski 1993). Systems could prompt users to summarize or reduce retrievals by criteria that do not necessarily touch upon their subject matter, e.g., date of publication, language, bibliographic format, library branch. Limiting results by format may not substantially reduce retrievals if the format sought by users is the predominant format in the library's bibliographic database. At academic institutions, reducing retrievals by branch libraries could touch upon the subject matter of retrievals because many institutions have branch libraries corresponding to academic disciplines, e.g., art library, chemistry library, engineering library, physics library.

Library classifications may hold considerable promise for summarizing the results of high-posted searches in terms of their subject matter. Systems could use broad ranges of classification numbers to consolidate retrievals and captions from the classification schedules to summarize the subjects of consolidated retrievals. When users selected a consolidated set of retrievals, online systems would respond with a display of subdivided forms using the available three or four types of coded subdivisions, i.e., topical, period, geographic, and, possibly, form. Depending on the number of retrievals, captions used to represent broad ranges of classification numbers could be taken from the almost two dozen alphabetical characters (A–Z) of the Library of Congress Classification (LCC), the ten numerical characters (0–9) of the Dewey Decimal Classification (DDC), or from smaller ranges within these larger ranges.

7.2Methodology

This chapter explores the use of library classifications for consolidating and summarizing large numbers of retrievals. We chose five moderately high-posted subjects entered by SULIRS, ORION, LS/2000, or MIRLYN users: (1) acid rain, (2) costa rica, (3) greek sculpture, (4) pornography, and (5) racism. All five queries were exact matches of controlled vocabulary terms. We searched these subjects in two sizable online catalogs: (1) Duke University's online catalog using its subject heading search (s=) that listed retrieved titles for subdivided and unsubdivided forms of the matched subject heading, and (2) The University of Michigan's MIRLYN online catalog using its subject heading search (s=) that listed retrieved titles for subdivided and unsubdivided forms of the matched subject heading. Retrievals in the Duke and Michigan catalogs were classified in the Dewey Decimal and Library of Congress Classifications, respectively.

We consolidated retrievals in classification number order and summarized large numbers of retrievals using captions from classification summaries and outlines. Retrievals for these subjects resulted in between one hundred and six hundred retrievals. We would have liked to have chosen subjects that were posted with thousands of retrievals. Unfortunately, the manual process we enlisted to search and consolidate retrievals was tremendously time consuming, complicated, and prone to human error to handle more than five or six hundred retrievals per subject. In the future, researchers should use computerized techniques to download thousands of retrievals and consolidate them according to the classification numbers and subject headings in retrieved records.

Retrievals were consolidated into categories that were described by classification captions. We examined two levels of consolidation using classification. The first level consolidated retrievals according to broad disciplines or subjects. In DDC, these broad disciplines were based on ten divisions (0-9). In LCC, these broad subjects would be based on the almost two dozen general classes (A-Z). When users selected a discipline or subject, systems would respond with second-level disciplines or subjects. In DDC, second-level disciplines would be based on the hundred divisions (00–99). In LCC, these broad subjects were based on the many of subjects represented by two alphabetical characters or classification-number ranges (for subjects that do not feature divisions using two alphabetical characters, i.e., E's, F's, Z's). When users select a second-level discipline or subject, systems would respond with alphabetical lists of subject headings that would be limited to the subject headings assigned to bibliographic records in the selected second-level discipline or subject. Selection of listed subject headings would result in the retrieval of titles assigned the headings. (If first- or second-level disciplines or subjects were not heavily posted, systems would display titles immediately rather than require users to proceed through intermediary displays of caption lists or subject heading lists.)

We also compared retrievals from searching the Dewey Decimal and Library of Congress Classification with retrievals from searching the two library catalogs to determine the extent of overlap between the two. Searching the DDC was easy using OCLC's Electronic DDC (EDDC). Searching LCC was much more difficult because the entire classification was not in machine-readable form. We manually entered LCC outlines into a FoxPro database and searched the database for the five selected topics. Since these topics retrieved bibliographic records in MIRLYN that were scattered into many different classification numbers, we checked the LCC schedule terminology of those that retrieved more than 2.5% of retrieved titles.

7.3 Summarization Using DDC

7.3.1 Acid Rain

Table 7.1 summarizes retrievals for the subject "acid rain" in Duke University's online catalog. (Appendix A lists all retrievals for this subject.)

Table 7.1. DDC Retrievals for "acid rain"

Summary category	Number
Number of titles retrieved	136
Number of unique classification numbers	37
Average no. of retrievals per number	3.7
Number of unique three-digit classification numbers	17
Frequently-occurring classification numbers, postings (in parentheses), and percentages of retrievals	363.7386 (35), 25.7% 363.7392 (20), 14.7% 363.7394 (12), 8.8%

The query "acid rain" retrieved 136 titles assigned the unsubdivided subject heading "Acid rain" and subdivided forms of this heading. Duke retrievals were distributed into 37 different classification numbers. There were 17 unique classification numbers beginning with the same three-digit number. The most common number was 363.7386 which accounted for a little over a quarter of titles. Overall, three classification numbers accounted for a little under 50% of retrieved titles.

Table 7.2 enlists first-level DDC captions to consolidate the 136 retrieved titles in Duke's online catalog on the subject of "acid rain."

Table 7.2. First-level DDC Captions for "acid rain"

DDC number	Number of titles	Classification captions
0	9	Generalities
3	96	Social sciences
5	8	Natural sciences and mathematics
6	21	Technology (Applied sciences)
7	1	The arts
9	1	Geography and history
0–9	136	Total

The majority of retrievals were found in the social sciences section of the Dewey Decimal Classification. If users selected the "Social sciences" caption, systems would summarize retrievals using second-level captions. Table 7.3 shows a summary based on the latter.

Table 7.3. Second-level DDC Captions for "acid rain" in the Social Sciences

DDC no.	No. of titles	Classification captions
32	1	Political science
34	10	Law
35	1	Public administration
36	84	Social services; association
30–39	96	Total

If users selected captions other than the "Social sciences" and "Social services" captions in Tables 7.2 and 7.3, systems would display retrieved titles because there were a manageable number of titles under each discipline. If they selected the "Social services" caption in Table 7.3, systems could summarize retrievals by displaying an alphabetical list of "Acid rain" subject headings assigned to these titles. Presumably, such a list would contain the unsubdivided subject heading "Acid rain" and several subdivided forms of this heading.

Keyword searches for "acid rain" in the Electronic DDC (EDDC) resulted in the six retrievals summarized in Table 7.4. On the far left is a column of DDC numbers in retrieved records. The “function” column designates the function of retrieved DDC terms and phrases. The “captions” column lists the captions for listed classification numbers. The number and percentage of titles that bear “Acid rain” subject headings are given on the far right.

Table 7.4. EDDC Retrievals for “acid rain”

DDC no.	Function of matched EDDC term(s)	Classification captions	No./% of titles	
341.7623	Index entry	Pollution control	4	2.9
344.04634	Index entry, Schedule including note	Control of the pollution of specific environments	0	0.0
363.73	Related DDC term	Pollution	0	0.0
363.7386	Index entry, common subject heading	Acid precipitation	35	25.7
551.5771	Index entry	Properties	0	0.0
628.532	Index entry	By products of combustion	4	2.9
Total	N/A	N/A	43	31.5

“Acid rain” did not occur in a caption. It occurred in five Relative Index entries, one including note, one common subject heading, and one related DDC term. (In the EDDC, common subject headings were created by processing a bibliographic database against the DDC schedules class-number database and identifying frequently-occurring subject headings for DDC class numbers listed in the schedules.)

The “titles” column in Table 7.4 refers to the number of titles that Duke searchers would retrieve in a class number search for this class number bearing “Acid rain” subject headings. The class number “363.7386” would retrieve about a quarter of the titles assigned the subject heading “Acid rain.” No titles would be retrieved in Duke’s catalog by the exact class numbers “344.04634” or “363.73” that bear the subject heading “Acid rain.” Searches for truncated forms of the former number would retrieve four titles. Searches of truncated forms of the latter number would retrieve 83 titles — almost two-thirds of the titles assigned “Acid rain” subject headings in Duke’s catalog. Exact and truncated forms of the class number “551.5771” retrieved no titles. Overall, the EDDC retrieved 31.5% of titles based on exact matches of retrieved classification numbers. If retrieval was based on truncated class numbers, the EDDC would retrieve 69.9% of titles that were assigned “Acid rain” subject headings. In the case of “acid rain,” searches of the EDDC would lead users to one moderately-posted classification number for this topic.

7.3.2 Costa Rica

Table 7.5 summarizes retrievals for the subject “costa rica” in Duke University’s online catalog. (Appendix B lists all retrievals for this subject.)

Table 7.5. DDC Retrievals for "costa rica

Summary category	Number
Number of titles retrieved	477
Number of unique classification numbers	142
Average no. of retrievals per number	3.4
Number of unique three-digit classification numbers	58
Frequently-occurring classification numbers, postings (in parentheses), and percentages of retrievals	972.86 (78), 16.4% 330.97286 (51), 10.7% 317.286 (16), 3.4% 972.8605 (16), 3.4% 972.86052 (16), 3.4%

The query "costa rica" retrieved 477 titles assigned the unsubdivided subject heading "Costa Rica" and subdivided forms of this heading. Duke retrievals were scattered into 142 different classification numbers. There were 58 unique classification numbers beginning with the same three-digit number. The most common number was 972.86; however, it accounted for only 16.4% of titles. Overall, the top three ranked classification numbers in terms of numbers of retrieved titles accounted for 37.3% of retrieved titles.

Table 7.6 enlists first-level DDC captions to consolidate the 477 retrieved titles in Duke's online catalog on the subject of "costa rica."

Table 7.6. First-level DDC Captions for "costa rica"

DDC number	Number of titles	Classification captions
0	9	Generalities
2	12	Religion
3	251	Social sciences
5	2	Natural sciences and mathematics
6	3	Technology (Applied sciences)
7	2	The arts
8	3	Literature and rhetoric
9	195	Geography and history
0–9	477	Total

Large numbers of retrievals were split between the social sciences (300s) and geography and history (900s) sections of the Dewey Decimal Classification. If users selected the captions representing these two sections, systems would summarize retrievals using second-level captions. Tables 7.7 and 7.8 show summaries based on second-level captions.

Table 7.7. Second-level DDC Captions for “costa rica” in the Social Sciences

DDC no.	No. of titles	Classification captions
30	25	Social sciences
31	29	General statistics
32	47	Political science
33	101	Economics
34	16	Law
35	22	Public administration
36	4	Social services; association
37	2	Education
38	4	Commerce, communications, transportation
39	1	Customs, etiquette, folklore
30–39	251	Total

Table 7.8. Second-level DDC Captions for “costa rica” in Geography and History

DDC no.	No. of titles	Classification captions
91	16	Geography and travel
92	6	Biography, genealogy, insignia
96	1	General history of Africa
97	170	General history of North America
98	2	General history of South America
90–99	195	Total

Except for the captions “Economics” and “General history of North America,” retrievals in Tables 7.7 and 7.8 were manageable for display. If users selected these two high-posted captions, systems could summarize retrievals by displaying an alphabetical list of “Costa Rica” subject headings assigned to these titles.

Keyword searches for “costa rica” in the Electronic DDC (EDDC) resulted in the twelve retrievals summarized in Table 7.9.

Table 7.9. EDDC Retrievals for “costa rica”

DDC no.	Function of matched EDDC term(s)	Classification captions	No./% of titles	
352.0073	Example note	Intermediate levels	0	0.0

819(.71–.79)	Add note	Specific countries	0	0.0
868.(99221–99227)	Add note	Specific countries	0	0.0
972.86	Schedule caption, Index entry	Costa Rica	78	16.4
–7286	Table 2 caption	Costa Rica	N/A	N/A
–72861	Index entry reference to Table 2	Limon Province	N/A	N/A
–72862	Index entry reference to Table 2	Cartago Province	N/A	N/A
–72863	Index entry reference to Table 2	San Jose Province	N/A	N/A
–72864	Index entry reference to Table 2	Heredia Province	N/A	N/A
–72865	Index entry reference to Table 2	Alajuela Province	N/A	N/A
–72866	Index entry reference to Table 2	Guanacaste Province	N/A	N/A
–72867	Index entry reference to Table 2	Puntarenas Province	N/A	N/A

Only one of twelve EDDC retrievals referred to a Schedules classification number that was posted in the Duke online catalog with bibliographic records bearing subdivided or unsubdivided forms of the subject heading “Costa Rica.” This retrieval referred to classification number “972.86” which was the highest posted classification number for this topic (Table 7.5). This exact number retrieved 16.4% of the titles bearing “Costa Rica” subject headings. Truncated numbers retrieved one-third of the titles bearing “Costa Rica” subject headings.

One of the twelve EDDC retrievals used the phrase “Costa Rica” in an example note on intermediate jurisdictional levels. The note reads “Examples: counties, districts, departments, arrondissements, Landkreise; provinces in certain jurisdictions, e.g., Costa Rica.” Materials on intermediate jurisdictional levels in many countries (including Costa Rica) were classed here; the phrase that refers to provinces in Costa Rica used “Costa Rica” as an example of a country with such a jurisdictional unit.

Eight table references were retrieved. The adoption of a tagging scheme for identifying the individual components of synthesized classification numbers would enable systems to use table references from the EDDC to produce retrievals. Several years ago, Arnold Wajenberg (1983) recommended a suitable approach to tagging these components but his ideas have not been adopted by the library community. In the absence of the adoption of a tagging scheme such as Wajenberg’s, it would be difficult for a system to parse synthesized classification numbers into individual components, and, thus, such numbers would not be useful retrievals.

In the case of “costa rica,” searches of the EDDC would lead users to one rather low-posted classification number for this topic.

7.3.3 Greek Sculpture

Table 7.10 summarizes retrievals for the subject “greek sculpture” in Duke University’s online catalog. (Appendix D lists all retrievals for this subject.)

Table 7.10. DDC Retrievals for “greek sculpture”

Summary category	Number
Number of titles retrieved	285
Number of unique classification numbers	55
Average no. of retrievals per number	5.2
Number of unique three-digit classification numbers	24
Top three frequently-occurring classification numbers, postings (in parentheses), and percentages of retrievals	733.3 (121), 42.5% 733 (71), 24.9% 913.38 (9), 3.2%

The query “greek sculpture” retrieved 285 titles assigned the unsubdivided subject heading “Greek sculpture” and subdivided forms of this heading. Duke retrievals were distributed into 55 different classification numbers and resulted in an average of 5.2 titles per unique classification number. There were 24 unique classification numbers beginning with the same three-digit number. The most common numbers were “733.3” and “733.” Together, these two numbers accounted for two-thirds of titles.

Table 7.11 enlists first-level DDC captions to consolidate the 285 retrieved titles in Duke’s online catalog on the subject of “greek sculpture.”

Table 7.11. First-level DDC Captions for “greek sculpture”

DDC number	Number of titles	Classification captions
0	3	Bibliography
7	239	The arts
8	3	Literature and rhetoric
9	40	Geography and history
0–9	285	Total

Only four main DDC classes contained titles assigned the subject heading “Greek sculpture.” If users selected classes 0, 8, or 9, systems would respond by displaying titles. If they selected class 7 in which the vast majority of numbers of retrievals were classed, systems would summarize retrievals using second-level captions. Table 7.12 shows a summary based on second-level captions.

Table 7.12. Second-level DDC Captions for "greek sculpture"

DDC number	Number of titles	Classification captions
70	18	The arts
72	7	Architecture
73	214	Plastic arts Sculpture
70–79	239	Total

Almost 90% of retrievals were classed in class "73" for plastic arts and sculpture. If users selected this class, systems would respond with unsubdivided and subdivided forms of the subject heading "Greek sculpture" to summarize the many retrievals in this class.

Keyword searches for "greek sculpture" in the Electronic DDC (EDDC) resulted in one retrieved Schedules caption: "733.3 Greek (Hellenic) sculpture." This exact number retrieved 42.5% of titles bearing the subject heading "Greek sculpture;" truncated forms of this number retrieved 46.3% of titles bearing this heading.

7.3.4 Pornography

Table 7.13 summarizes retrievals for the subject "pornography" in Duke University's online catalog. (Appendix D lists all retrievals for this subject.)

Table 7.13. DDC Retrievals for "pornography"

Summary category	Number
Number of titles retrieved	111
Number of unique classification numbers	43
Average no. of retrievals per number	2.6
Number of unique three-digit classification numbers	26
Frequently-occurring classification numbers, postings (in parentheses), and percentages of retrievals	363.47 (44), 39.6% 016.36347 (6), 5.4% 363.470973 (5), 4.5%

The query "pornography" retrieved 111 titles assigned the unsubdivided subject heading "Pornography" and subdivided forms of this heading. Duke retrievals were scattered into 43 different classification numbers; however, almost 40% of titles were given the class number, 363.47. No other frequently-occurring number came close in terms of numbers of retrievals. Overall, the top three ranked classification numbers in terms of numbers of retrieved titles accounted for a little under 50% of retrieved titles.

Table 7.14 enlists first-level DDC captions to consolidate the 111 retrieved titles in Duke's online catalog on the subject of "pornography."

Table 7.14. First-level DDC Captions for "pornography"

DDC number	Number of titles	Classification captions
0	10	Generalities
1	3	Philosophy and psychology
2	5	Religion
3	82	Social sciences
4	1	Language
7	3	The arts
8	3	Literature and rhetoric
9	4	Geography and history
0–9	111	Total

All but two main DDC classes contained titles assigned the subject heading "Pornography." Except for class 3, the numbers of retrievals under each class were manageable for displays of retrieved titles. If users selected the "Social sciences" class, systems would summarize retrievals using second-level captions. Table 7.15 shows a summary based on second-level captions.

Table 7.15. Second-level DDC Captions for "pornography"

DDC number	Number of titles	Classification captions
30	14	Social sciences
34	5	Law
36	63	Social services; association
30–39	82	Total

Over three-quarters of retrievals were classed in class "36" for social services. If users selected this class, systems would respond with unsubdivided and subdivided forms of the subject heading "Pornography" to summarize the moderate number of retrievals in this class.

Keyword searches for "pornography" in the Electronic DDC (EDDC) resulted in the ten retrievals summarized in Table 7.16.

Table 7.16. EDDC Retrievals for "pornography"

DDC no.	Function of matched EDDC term(s)	Classification captions	No./% of titles	
306.77	See also note	Sexual practices	3	2.7

341.77	Including note	International criminal law	0	0.0
342.0853	See also note	Promulgation of information and opinion	0	0.0
344.0547	Index entry for built number	(Closest Schedules number/ caption: 344.05/ Police services)	1	0.9
345.0274	Index entry for built number	(Closest Schedules number/ caption: 345.02/Crimes (Offenses))	0	0.0
363.47	Caption, class elsewhere note, index entry, common subject heading	Obscenity and pornography	44	39.6
364.174	Caption, class elsewhere note, index entry	Obscenity and pornography	3	2.7
704.9428	Including note	Erotica	2	1.8
808.803538	Index entry for built number	(Closest Schedules number/ caption: 808.8/Collections of literary texts from more than one literature)	0	0.0
809.933538	Index entry for built number	(Closest Schedules number/ caption: 809.933/Dealing with specific themes and subjects)	1	0.9
–0803538	Index entry reference to Table 3B	(Closest Tables number/ caption: –08/Collections of literary texts in more than one form)	N/A	N/A
–093538	Index entry reference to Table 3B	(Closest Tables number/ caption: –09/History, description, critical appraisal of works in more than one form)	N/A	N/A
Total	N/A	N/A	54	48.7

Captions, index entries, common subject headings, and notes referred to “pornography.” Of the two captions that referred to this term, the class number for one caption was the same as the call number for the highest posted class number in online searches of Duke’s catalog, viz. 363.47. This number retrieved about 40% of the titles assigned unsubdivided and subdivided forms of the subject heading “Pornography.” EDDC retrievals included eight Index entries. Four entries referred users to built numbers based on the Schedules and two entries referred them to built numbers based on Table 3B; Table 7.16 lists closest Schedules or Tables numbers and captions in the “DDC no.”column. Overall, EDDC searches for “pornography” led to exact class numbers that retrieved 48.7% of the titles assigned “Pornography” subject headings and to truncated numbers that retrieved 62.2% of titles assigned these headings.

7.3.5 Racism

Table 7.17 summarizes retrievals for the subject “racism” in Duke University’s online catalog. (Appendix E lists all retrievals for this subject.)

Table 7.17. DDC Retrievals for “racism”

Summary category	Number
Number of titles retrieved	347
Number of unique classification numbers	163
Average no. of retrievals per number	2.1
Number of unique three-digit classification numbers	60
Frequently-occurring classification numbers, postings (in parentheses), and percentages of retrievals	305.800973 (41), 11.8% 305.8 (40), 11.5% 305.896073 (25), 7.2%

A total of 347 retrievals were scattered into 163 different classification numbers. On the average, 2.1 titles per classification number were retrieved; this average number was the lowest such number for the five subjects examined in this section. There were 60 unique classification numbers beginning with the same three-digit number. The most common number was a synthesized number — 305.800973 — made from the Schedules number 305.8 (“Racial, ethnic, national groups”), Table 1 number “009” (“Historical, geographical, and persons treatment,” and Table 2 number “73” (“United States”). The synthesized number 305.800973 accounted for only 11.8% of retrieved titles. The Schedules number, 305.8, was very close behind; it accounted for 11.5% of retrieved titles. Overall, the top three-ranked classification numbers in terms of numbers of retrieved titles accounted for 30.5% of retrieved titles.

Table 7.18 enlists first-level DDC captions to consolidate the 347 retrieved titles in Duke’s online catalog on the subject of “racism.”

Table 7.18. First-level DDC Captions for “racism”

DDC number	Number of titles	Classification captions
0	3	Generalities
1	6	Philosophy and psychology
2	16	Religion
3	282	Social sciences
5	2	Natural sciences and mathematics
6	1	Technology (Applied sciences)
7	1	The arts

8	3	Literature and rhetoric
9	33	Geography and history
0–9	347	Total

“Racism” was addressed by nine of the ten DDC major classes. Most retrievals were classed in the “Social sciences” class. If users selected a class other than class 3, “Social sciences,” from Table 7.18, systems would respond by displaying titles because there was a manageable number of titles to browse. If they selected class 3 in which the vast majority of numbers of retrievals were classed, systems would summarize retrievals using second-level captions. Table 7.19 shows a summary based on second-level captions for class 3.

Table 7.19. Second-level DDC Captions for “racism”

DDC number	Number of titles	Classification captions
30	215	Social sciences
32	35	Political science
33	9	Economics
34	7	Law
35	1	Public administration
36	6	Social services; association
37	9	Education
30–39	282	Total

Table 7.19 shows manageable numbers of retrievals in six of seven classes. Most titles (76.2%) were assigned to class “30” for “Social sciences.” If users selected this class, systems would respond with unsubdivided and subdivided forms of the subject heading “Racism” to summarize the large number of retrievals in this class.

Keyword searches for “racism” in the Electronic DDC (EDDC) resulted in the six retrievals summarized in Table 7.20.

Table 7.20. EDDC Retrievals for “racism”

DDC no.	Function of matched EDDC term(s)	Classification captions	No./% of titles	
172–179	Centered heading	Applied ethics (Social ethics)	0	0.0
177.5	Schedules caption	Slavery and discriminatory practices	0	0.0
261.8348	Index entry for built number	(Closest Schedules number/ caption: 261.834/Social structure)	5	1.4

291.178348	Index entry for built number	(Closest Schedules number/ caption: 291.17834/Social structure)	0	0.0
305.8	Index entry	Racial, ethnic, national groups	40	11.5
320.56	Schedules caption	Racism	5	1.4
Total	N/A	N/A	50	14.3

The term “racism” occurred in only one DDC caption (320.56); only five titles were assigned its classification number in Duke’s bibliographic database. The Schedules class number 305.8 accounted for 11.5% of titles from Duke’s database. Only one of the remaining four EDDC retrievals cited a classification number (261.8348) in which titles were classed in the Duke database. There were no EDDC retrievals for the synthesized class number “305.800973” which retrieved the largest number (41) of retrievals in a search of the Duke online catalog. If retrieval was based on exact matches of retrieved classification numbers, the EDDC retrieved 14.3% of the titles that were assigned “Racism” subject headings; if retrieval was based on truncated classification numbers, the EDDC retrieved 42.9% of the titles that were assigned these headings

7.4Summarization Using LCC

7.4.1 Acid Rain

Table 7.21 summarizes retrievals for the subject “acid rain” in the University of Michigan’s MIRLYN online catalog. (Appendix F lists all retrievals for this subject.)

Table 7.21. LCC Retrievals for “acid rain”

Summary category	Number
Number of titles retrieved	141
Number of unique classification numbers	30
Average no. of retrievals per number	4.7
Number of unique classification numbers (up to the decimal point)	25
Frequently-occurring classification numbers, postings (in parentheses), and percentages of retrievals	TD196.A25 (73), 51.8% QH545.A17 (20), 14.2% HD9685.C2 (4), 2.8% Z5862.2.A26 (4), 2.8%

The query “acid rain” retrieved 141 titles assigned the unsubdivided subject heading “Acid rain” and subdivided forms of this heading. MIRLYN retrievals were distributed into 30 different Library of Congress Classification numbers. There were 25 unique classification numbers beginning with the same number disregarding alphanumeric characters beyond the decimal point. The most common number was TD196.A25 which retrieved 51.8% of the titles bearing "Acid rain" subject headings. In the analysis of summarization using the DDC, no single number accounted for more than 42.5% of retrievals ("greek sculpture," Table 7.10). The class number "QH545.A17" accounted for almost 14% of retrieved titles. No other classification numbers came close to the magnitude of these two frequently-occurring classification numbers.

Table 7.22 enlists first-level DDC captions to consolidate the 141 retrieved titles in Michigan's online catalog on the subject of "acid rain."

Table 7.22. First-level LCC Captions for "acid rain"

LCC number	Number of titles	Classification captions
H	6	Social sciences
J	1	Political science
K	3	Law
Q	32	Science
S	5	Agriculture
T	88	Technology
Z	6	Bibliography and Library Science
A–Z	141	Total

The majority of retrievals were found in the Technology section of the Library of Congress Classification. This was different from the result for DDC retrievals in which most retrievals were concentrated in the Social sciences class (Table 7.2). Manageable numbers of retrievals occurred in all classes but the Technology class. If users selected low-posted classes, systems would display brief-titles lists. If they selected the Technology class, systems would summarize retrievals using second-level captions. Table 7.23 shows a summary based on second-level classes in the Technology class.

Table 7.23. Second-level LCC Captions for "acid rain" in Technology

LCC no.	No. of titles	Classification captions
TA	1	Technology (General)
TD	87	Environmental technology. Sanitary engineering.
TA–TX	88	Total

Titles were summarized by only two second-level LCC classes and all but one title was classed in "TD." The second-level display did not summarize retrievals with greater distinction than the original, first-level classes. If class number ranges from the classification outline were used instead of second-level displays, the 88 titles would still be summarized by two classes and almost all titles would occur in one of the two classes: (1) one title in the range TA164–TA1280 for "Engineering — general; Civil engineering — general," and (2) 87 titles in the range TD159–TD949 for "Environmental technology; Sanitary engineering." Another possibility for consolidating retrievals would be third-level displays. Table 7.24 shows a summary based on third-level classes.

Table 7.24. Third-level LCC Captions for "acid rain" in Technology

LCC no.	No. of titles	Classification captions
TA418	1	General works [Elastic properties and tests]
TD195	10	Special industries, facilities, activities, etc.
TD196	73	Other environmental pollutants, A–Z
TD427	1	Special pollutants and organisms, A–Z
TD883	1	General works [Air pollution and its control]
TD885	2	General works [Gases. Flue gases]
TA–TX	88	Total

The third-level display distributed titles into six classes. Although the majority (83.0%) of titles occurred in a single class, two classes featured more than five titles. The search for "acid rain" resulted in a little less than 150 titles. When searches result in thousands of retrievals, third-level displays are likely to become extremely long, and, possibly, as difficult to browse as lengthy subject heading displays. The same, awkward wording was used for three of the six LCC captions, viz. "General works." Placed in brackets and accompanying these three phrases in Table 7.24 are captions at the next highest level of the classification which give "General works" captions context.

Keyword searches for "acid rain" in machine-readable LCC outline records resulted in no retrievals. We checked the terminology of LCC schedule captions for class numbers for which MIRLYN searches retrieved over 2.5% of the total number of retrievals in searches for the subject heading "Acid rain." Table 7.25 summarizes the results.

Table 7.25. LCC Schedule Retrievals for "acid rain"

LCC no.	Schedule captions	No./% of titles	
HD9685	General works [Electric utilities and industries]	4	2.8
QH545.A17	Acid rain. Acid precipitation. Acid deposition	20	14.2
TD196.A25	Acid rain [Other environmental pollutants, A–Z; Special environmental pollutants]	73	51.8

Z5862.2	Acid rain [Pollution and pollution control; Environment (general and human). Human ecology; Subject bibliography]	6	4.3
Total	N/A	103	73.1

In Table 7.26, we have enclosed captions at higher level(s) of the Library of Congress Classification in brackets to give the listed captions context. Together, the four captions accounted for 73.1% of the titles assigned "Acid rain" subject headings. Three of the four captions enlisted the same terminology as the subject heading "Acid rain."

7.4.2 Costa Rica

Table 7.26 summarizes retrievals for the subject "costa rica" in Michigan's MIRLYN online catalog. (Appendix G lists all retrievals for this subject.)

Table 7.26. LCC Retrievals for "costa rica

Summary category	Number
Number of titles retrieved	344
Number of unique classification numbers	109
Average no. of retrievals per number	3.2
Number of unique classification numbers (up to the decimal point)	94
Frequently-occurring classification numbers, postings (in parentheses), and percentages of retrievals	HC143 (39), 11.3% F1546 (34), 9.9% F1549.B7 (33), 9.6%

The query "costa rica" retrieved 344 titles assigned the unsubdivided subject heading "Costa Rica" and subdivided forms of this heading. MIRLYN retrievals were scattered into 109 different classification numbers. There were almost as many unique classification numbers beginning with the same number up to the decimal point (94) as unique classification numbers (109). The most common number was "HC143." It accounted for only 11.3% of titles. Overall, the top three ranked classification numbers in terms of numbers of retrieved titles accounted for 30.8% of retrieved titles.

Table 7.27 enlists first-level LCC captions to consolidate the 344 retrieved titles in the MIRLYN online catalog on the subject of "costa rica."

Table 7.27. First-level LCC Captions for "costa rica"

LCC number	Number of titles	Classification captions
A	9	General works
B	12	Philosophy, psychology, religion

C	2	Auxiliary sciences of history
D	1	World history
E	2	History of North and South America
F	177	History of North and South America
G	20	Geography, maps, anthropology, recreation
H	74	Social sciences
J	23	Political science
L	3	Education
P	2	Languages and literature
Q	6	Science
R	1	Medicine
S	1	Agriculture
Z	11	Bibliography and library science
A–Z	344	Total

Large numbers of retrievals were split between the “Social sciences” (H) and “History of North and South America” (F) sections of LCC. This split was the same as the split for retrievals in the DDC (Table 7.6). Numbers of retrievals in other classes were manageable; if users selected them, systems would show brief-titles lists. If users selected the captions representing the two high-posted classes, systems could summarize retrievals using second-level captions. Tables 7.28 and 7.29 show summaries based on second-level captions.

Table 7.28. Second-level LCC Captions for “costa rica” in History

LCC no.	No. of titles	Classification captions
F1401–1419	4	Latin America (General). Spanish America (General)
F1421–1440	4	Central America
F1521–1537	8	Nicaragua
F1541–1557	161	Costa Rica
F’s	177	Total

Table 7.29. Second-level LCC Captions for “costa rica” in the Social Sciences

LCC no.	No. of titles	Classification captions
H	3	Social sciences (General)
HA	7	Statistics
HB	3	Economic theory

HC	42	Economic history and conditions
HD	3	Economic history and conditions
HE	1	Transportation and communications
HF	1	Commerce
HG	2	Finance
HJ	3	Public finance
HM	1	Sociology (General)
HN	4	Social history and conditions. Social problems. Social reform
HQ	3	The family. Marriage. Woman
HX	1	International law
H–HX	74	Total

Since class F has no classes beginning with two-letter combinations, we used classification number ranges to represent second-level classes. Over 90% of retrievals were classed in the "Costa Rica" range, "F1541–F1557."

A total of thirteen classes summarized retrievals in the "H" ("Social sciences") class. One class ("HC") summarized 56.8% of retrievals in this class. If low-posted classes (i.e., classes with one or two retrievals) were combined into a single "Other social sciences" class, the number of classes would be reduced to nine classes.

Two classification captions (HC and HD) were worded exactly alike, viz. "Economic history and conditions." Retrievals could be combined into a single caption; however, some searches might have different results in which the separation between retrievals from the two classes might be helpful. To reduce end-user confusion, new name(s) should be given to one or both captions. The caption term "Woman" under HQ could be offensive to some end users. Second-level results in the social sciences demonstrate the need to review LCC schedule terminology to improve understanding and reduce awkward or potentially offensive terminology.

Keyword searches for "costa rica" in machine-readable LCC outline records resulted in two retrievals. Both captions had the same terminology, i.e, "Costa Rica." One caption covered the range F1541–F1557 in which 161 (46.8%) titles were also assigned the subject heading "Costa Rica." The other caption covered the range PQ7480–PQ7489 but only one title in this range was assigned a "Costa Rica" subject heading.

We checked the terminology of LCC schedule captions for class numbers for which MIRLYN searches retrieved over 2.5% of the total number of retrievals in searches for the subject heading "Costa Rica." Table 7.30 summarizes the results.

Table 7.30. LCC Schedule Retrievals for "costa rica"

LCC no.	Schedule captions	No./% of titles	
F1543	General works [Costa Rica]	14	4.1
F1546	General works [History; Costa Rica]	34	9.9
F1547	1502–1838	9	2.6
F1547.5	1838–1948	16	4.7
F1548	1948–1986	19	5.5
F1549.B7	Boundaries	33	9.6
G4860	Costa Rica [Central America; North America; America, Northern hemisphere; By region or country; Maps]	9	2.6
HC143	Costa Rica [Central America; North America; America, Northern hemisphere; By region or country; Economic history and conditions]	39	11.3
Total	N/A	173	50.3

In Table 7.31, we have enclosed captions at higher level(s) of the Library of Congress Classification in brackets to give the listed captions context. The context for the last two-listed captions was not evident until reaching the highest-level captions, i.e., "Maps" and "Economic history and conditions." Together, the six captions accounted for 50.3% of the titles assigned "Costa Rica" subject headings. Only two captions enlisted the same terminology as the subject heading "Acid rain."

7.4.3 Greek Sculpture

Table 7.31 summarizes retrievals for the subject "greek sculpture" in MIRLYN. (Appendix H lists all retrievals for this subject.)

Table 7.31. LCC Retrievals for "greek sculpture"

Summary category	Number
Number of titles retrieved	442
Number of unique classification numbers	93
Average no. of retrievals per number	4.8
Number of unique classification numbers (up to the decimal point)	89
Top three frequently-occurring classification numbers, postings (in parentheses), and percentages of retrievals	NB90 (98), 22.2% NB91 (49), 11.1% N13 (48), 10.9%

The query "greek sculpture" retrieved 442 titles assigned the unsubdivided subject heading

"Greek sculpture" and subdivided forms of this heading. MIRLYN retrievals were distributed into 93 different classification numbers and resulted in an average of 4.8 titles per unique classification number. There were almost as many unique classification numbers beginning with the same number up to the decimal point (89) as unique classification numbers (93). The most common numbers were NB90, NB91, and N13. Together, these three numbers accounted for 44.2% of titles.

Table 7.32 enlists first-level DDC captions to consolidate the 442 retrieved titles in MIRLYN on the subject of "greek sculpture."

Table 7.32. First-level DDC Captions for "greek sculpture"

LCC number	Number of titles	Classification captions
A	18	General works
B	1	Philosophy, psychology, and religion
C	5	Auxiliary sciences of history
D	43	World history
N	372	Fine Arts
P	1	Languages and literature
Q	2	Science
A–Z	442	Total

The majority (84.2%) of titles were classed in N, "Fine arts." This was the same class as the majority of DDC retrievals (Table 7.11). Seven main LCC classes contained titles assigned the subject heading "Greek sculpture." All but one class (N) yielded manageable numbers of titles, thus, systems would respond to users who selected these classes (A, B, C, D, P, and/or Q) by displaying titles. If they selected class N in which the vast majority (84.2%) of numbers of retrievals were classed, systems would summarize retrievals using second-level captions. Table 7.33 shows a summary based on second-level captions.

Table 7.33. Second-level DDC Captions for "greek sculpture" in the Fine Arts

LCC number	Number of titles	Classification captions
N	89	Visual arts
NA	7	Architecture
NB	273	Sculpture
ND	1	Painting
NK	2	Decorative arts. Applied arts. Decoration and ornament
N–NX	372	Total

Table 7.33 shows two high-posted classes: N ("Visual arts") and NB ("Sculpture"). If users selected these classes, systems would respond with unsubdivided and subdivided forms of the subject heading "Greek sculpture" to summarize the many retrievals.

Keyword searches for "greek sculpture" in machine-readable LCC outline records resulted in no retrievals. We checked the terminology of LCC schedule captions for class numbers for which MIRLYN searches retrieved over 2.5% of the total number of retrievals in searches for the subject heading "Greek sculpture." Table 7.34 summarizes the results.

Table 7.34. LCC Schedule Retrievals for "greek sculpture"

LCC no.	Schedule captions	No./% of titles	
DE2	Societies [The Mediterranean Region; The Greco-Roman world]	12	2.7
N13	French and Belgian [Societies; Visual arts]	48	10.9
NB87	Special collections [Classical; Ancient sculpture; History; Sculpture]	18	4.1
NB90	General works [Greek, Ancient sculpture, History; Sculpture]	98	22.2
NB91	Special localities (place of origin) [Greek, Ancient sculpture, History; Sculpture]	49	11.1
NB94	General special [Greek, Ancient sculpture, History; Sculpture]	26	5.9
Total	N/A	251	56.8

Almost all captions would probably not make sense to end users — "general works," "special collections," "special localities," "general special." The captions at higher level(s) of the Library of Congress Classification that are enclosed in brackets give the listed captions context. Together, the six captions accounted for over 50% of the titles assigned "Greek sculpture" subject headings. None of the captions enlisted the same terminology as the subject heading "Greek sculpture."

7.4.4 Pornography

Table 7.35 summarizes retrievals for the subject "pornography" in MIRLYN. (Appendix I lists all retrievals for this subject.)

Table 7.35. LCC Retrievals for "pornography"

Summary category	Number
Number of titles retrieved	194
Number of unique classification numbers	51

Average no. of retrievals per number	3.8
Number of unique classification numbers (up to the decimal point)	48
Frequently-occurring classification numbers, postings (in parentheses), and percentages of retrievals	HQ471 (94), 48.5% KF9444 (11), 5.7% HQ472 (10), 5.2%

The query “pornography” retrieved 194 titles assigned the unsubdivided subject heading “Pornography” and subdivided forms of this heading. MIRLYN retrievals were scattered into 51 different classification numbers; however, almost 50% of titles were given the same LC class number, HQ471. (Interestingly, almost 40% of titles were given the same DDC class number for this subject, Table 7.13). No other number came close in terms of numbers of LCC retrievals. (No other number came close in terms of DDC retrievals also, Table 7.13). Overall, the top three ranked classification numbers in terms of numbers of retrieved titles accounted for a little under 60% of retrieved titles.

Table 7.36 enlists first-level LCC captions to consolidate the 194 retrieved titles in MIRLYN on the subject of “pornography.”

Table 7.36. First-level LCC Captions for “pornography”

LCC number	Number of titles	Classification captions
B	2	Philosophy, psychology, and religion
H	141	Social sciences
J	7	Political science
K	17	Law
N	1	Fine arts
P	8	Languages and literature
R	2	Medicine
Z	16	Bibliography and library science
A–Z	194	Total

The majority (72.7%) of titles were classed in “Social sciences.” (“Social sciences” was the main class in which most DDC retrievals were classed, Table 7.14). Only the “Social sciences” class bears an unmanageable number of LCC retrievals. Table 7.37 shows a summary of “Social sciences” retrievals based on second-level captions.

Table 7.37. Second-level DDC Captions for "pornography" in the Social Sciences

LCC number	Number of titles	Classification captions
HE	1	Transportation and communications
HG	1	Finance
HM	1	Sociology (general)
HN	3	Social history and conditions. Social problems. Social reform
HQ	125	The family. Marriage. Woman
HV	10	Social pathology. Social and public welfare. Criminology
H–HX	141	Total

Second-level classes distributed 141 titles into six more specific classes. Unfortunately, almost 90% of retrievals occurred in a single class (HQ, "The family. Marriage. Woman"). If users selected this class, systems would respond with unsubdivided and subdivided forms of the subject heading "Pornography" to summarize the large number of retrievals in this class.

Keyword searches for "pornography" in machine-readable LCC outline records resulted in one retrieval — the range HQ450–HQ471 which was summarized by the caption "Erotica. Pornography." This range retrieved 54.6% of titles assigned "Pornography" subject headings in the MIRLYN database.

We checked the terminology of LCC schedule captions for class numbers for which MIRLYN searches retrieved over 2.5% of the total number of retrievals in searches for the subject heading "pornography." Table 7.38 summarizes the results.

Table 7.38. LCC Schedule Retrievals for "pornography"

LCC no.	Schedule captions	No./% of titles	
HQ471	General works [Pornography. Obscene literature; Erotica; The family. Marriage. Woman.]	94	48.5
HQ472	By region or country, A–Z [Pornography. Obscene literature; Erotica; The family. Marriage. Woman.]	10	5.2
HV6727	Obscene literature [Offenses against public morals; Crimes and offenses; Criminology]	11	5.7
KF9444	Obscenity [Criminal law; United States — general]	11	5.7
Z657	General works [Freedom of press. Censorship]	9	4.6
Total	N/A	135	69.7

The terminology and context were not clear for the captions "General works" and "By region or country, A–Z. " The captions at higher level(s) of the Library of Congress Classification that were enclosed in brackets gave these lower-level captions context. Together, the five captions accounted for over two-thirds of the titles assigned "Pornography" subject headings.

7.4.5 Racism

Table 7.39 summarizes retrievals for the subject "racism" in MIRLYN. (Appendix J lists all retrievals for this subject.)

Table 7.39. LCC Retrievals for "racism"

Summary category	Number
Number of titles retrieved	497
Number of unique classification numbers	236
Average no. of retrievals per number	2.1
Number of unique classification numbers (up to the decimal point)	196
Frequently-occurring classification numbers, postings (in parentheses), and percentages of retrievals	HT1521 (56), 11.3% DA125 (31), 6.2% E185.615 (22), 4.4%

A total of 497 retrievals were scattered into 236 different classification numbers. On the average, 2.1 titles per classification number were retrieved; this average number was the lowest such number for the five subjects examined in this section. (DDC retrievals for "racism" also recorded the lowest average number of titles for this subject, Table 7.17). There were almost as many unique classification numbers beginning with the same number up to the decimal point (196) as unique classification numbers (236). The most common numbers occurred in three different first-level class numbers: HT1521, DA125, and E185.615. Together, these three numbers accounted for 21.9% of titles.

Table 7.40 enlists first-level DDC captions to consolidate the 497 retrieved titles in Michigan's online catalog on the subject of "racism."

Table 7.40. First-level LCC Captions for "racism"

DDC number	Number of titles	Classification captions
A	1	General works
B	16	Philosophy, psychology, religion
C	3	Auxiliary sciences of history
D	83	World history
E	99	History of North and South America
F	27	History of North and South America

G	11	Geography, maps, anthropology, recreation
H	144	Social sciences
J	23	Political science
K	12	Law
L	44	Education
M	1	Music
N	1	Fine arts
P	15	Languages and literature
R	7	Medicine
Z	10	Bibliography and library science
A–Z	497	Total

“Racism” was addressed by sixteen LCC major classes. More than two dozen retrievals occurred in classes D, E, F, H, and L. Three classes resulted in unmanageable numbers of retrievals: D, E, and H. If users selected these classes, systems would summarize retrievals using second-level captions. Table 7.41 shows summaries based on second-level captions for class D, “World history.

Table 7.41. Second-level LCC Captions for “racism” in World History

LCC number	Number of titles	Classification captions
D	8	History (general)
DA	37	Great Britain
DB	1	Austria. Liechtenstein, Hungary. Czechoslovakia
DC	13	France
DD	8	Germany
DJ	2	Netherlands (Holland)
DP	2	Spain
DR	1	Balkan Peninsula
DS	7	Asia
DT	1	Africa
DU	3	Oceania (South Seas)
D–DX	83	Total

Second-level DDC captions for “racism” under “World history” distributed retrievals into manageable divisions based on geography.

Table 7.42 shows summaries based on second-level captions for class E, “History of North and

South America."

Table 7.42. Second-level LCC Captions for "racism" in the History of North and South America

LCC range	Number of titles	Classification captions
E11–E29	2	General
E171–E179.5	4	History. General
E184–E185.9782	82	Elements in the population
E415.6–E440.5	1	Middle 19th century
E441–E453	3	Slavery in the United States
E666–E670	1	Johnson's administration
E747–E748	1	Biography
E835–E839	5	Eisenhower's administration
E's	99	Total

Since class E had no classes beginning with two-letter combinations, we used classification number ranges to represent second-level classes. Manageable numbers of retrievals occurred in seven of eight classes. Most (82.8%) titles were classed in the range "E184–E185.9782" for "Elements in the population." If users selected this class, systems would respond with unsubdivided and subdivided forms of the subject heading "Racism" to summarize the large number of retrievals in this class.

Table 7.43 shows summaries based on second-level captions for class H, "Social sciences."

Table 7.43. Second-level LCC Captions for "racism" in the Social sciences

LCC number	Number of titles	Classification captions
H	2	Social sciences (general)
HC	2	Economic history and conditions
HD	7	Economic history and conditions
HM	10	Sociology
HN	9	Social history and conditions. Social problems. Social reform
HQ	17	The family. Marriage, Woman
HS	13	Societies: Secret, benevolent, etc.
HT	67	Communities. Classes. Races
HV	7	Social pathology. Social and public welfare
HX	10	Socialism. Communism. Anarchism

H–HX	144	Total

A total of 144 titles were distributed into ten second-level "Social sciences" classes. Retrievals were manageable except for class HT. Some classification captions in Tables 7.41–7.43 might need to be reworded to reduce duplication, i.e., two captions with the phrase "Economic history and conditions," to make sure end users understand them, i.e., captions bearing the word "general, " "Societies: Secret, benevolent, etc.," "Elements of the population," or are not offended by them, i.e., caption bearing the word "Woman."

Keyword searches for "racism" in machine-readable LCC outline records resulted in no retrievals. We checked the terminology of LCC schedule captions for class numbers for which MIRLYN searches retrieved over 2.5% of the total number of retrievals in searches for the subject heading "Racism." Table 7.44 summarizes the results.

Table 7.44 LCC Schedule Retrievals for "racism"

LCC no.	Schedule captions	No./% of titles	
DA125	Elements in the population [Ethnography; General special; History; England; Great Britain]	31	6.2
E184.A1	General works [Elements, A–Z; Elements in the population; United States]	19	3.8
E185.61	Race relations [1877–1964; History (By period); Afro-Americans; Elements in the population; United States]	12	2.4
E185.615	1964– [History; Afro-Americans; Elements in the population; United States]	22	4.4
HT1521	General works [Races; Communities; Classes; Races]	56	11.3
Total	N/A	140	28.1

The terminology and context were not clear for the captions "General works" and "Elements of the population." The captions at higher level(s) of the Library of Congress Classification that were enclosed in brackets gave these lower-level captions context; however, captions like "Elements, A–Z" and "General special" would probably not be understandable to end users. Together, the five captions accounted for 28.1% of the titles assigned "Racism" subject headings. None of the lower- or higher-level captions enlisted the same terminology as the subject heading "Racism."

7.5Using Classification to Summarize Large Retrievals

We searched five moderately high posted subjects to explore how the Dewey Decimal and Library of Congress Classifications would summarize large retrievals. The subjects were "acid rain," "costa rica," "greek sculpture," "pornography," and "racism."

Numbers of titles retrieved for these five subjects in Duke's online catalog ranged from a low of 111 titles ("pornography") to a high of 477 titles ("costa rica"); numbers of titles retrieved for these five subjects in Michigan's online catalog ranged from a low of 141 titles ("acid rain") to a high of 497 titles ("racism"). The average number of retrievals per unique, retrieved DDC classification number ranged from a low of 2.1 titles ("racism") to a high of 5.2 titles ("greek sculpture") and per LCC classification number ranged from a low of 2.1 titles ("racism") to a high of 4.8 titles ("greek sculpture"). The lowest percentages of titles retrieved for the most common DDC number and LCC number for a particular subject were 11.8% for the synthesized DDC number "305.800973" under "racism" and 11.3% for "HC143" and "HT1521" under "costa rica" and "racism," respectively. The highest percentages of titles retrieved for the most common DDC number and LCC number for a particular subject were 42.5% for the DDC schedules number "733.3" under "greek sculpture" and 51.8% for the LCC number "TD196.A25" under "acid rain." Although our analysis of subject heading and classification number retrievals was far from comprehensive, the analysis resulted in several generalizations.

Generalizations about retrievals:

- Retrievals were distributed into many unique classification numbers across many broad disciplines.
- Retrievals were likely to result in one common classification number; however, the percentage of retrievals for such numbers could vary considerably, i.e., between ten and fifty percent of the titles assigned the subject heading.
- Retrievals were not likely to result in more than two common classification numbers that, together, retrieved more than half of the titles assigned the subject heading.
- Retrievals sometimes resulted in common classification numbers that exactly matched Schedules classification numbers (DDC only).
- Retrievals sometimes resulted in common classification numbers that matched synthesized classification numbers (DDC only).

When we applied first-level DDC captions to summarize retrievals, a total of only six captions were connected with an unmanageable number of retrievals. One subject featured two first-level captions with an unmanageable number of retrievals ("Costa Rica") and the remaining four subjects featured a single first-level caption with an unmanageable number of retrievals. When we applied first-level LCC captions to summarize retrievals, a total of only ten captions were connected with an unmanageable number of retrievals. Two subjects featured two first-level captions with an unmanageable number of retrievals ("costa rica" and "greek sculpture") and one subject featured three first-level captions with an unmanageable number of retrievals ("racism").

When users select captions with unmanageable numbers of retrievals, systems should summarize retrievals using the subject headings that were the impetus for the original search of the catalog. We would have liked to have shown such a display of subject headings but the catalogs we searched did not feature subject heading searches that could be limited to specific classification numbers. Future explorations of the use of classification to summarize retrievals should experiment with such displays. They should also focus on very high-posted subjects — subjects that retrieve several thousands of retrievals — because, in addition to using classification for summarization, such results are likely to also require summarization of subject headings.

Generalizations about library classification classes:

- The majority of first- and second-level classes were assigned to manageable numbers of titles.
- Retrievals were likely to result in one first-level class that retrieved between about half to three-quarters of the titles assigned the subject heading.
- Retrievals were not likely to result in two or more first-level classes that, on their own, were assigned to sizable numbers of titles.
- Retrievals were likely to result in one second-level class that retrieved between about half to three-quarters of the titles assigned the subject heading.
- Retrievals were not likely to result in two or more second-level classes that, on their own, were assigned to sizable numbers of titles.

For the most part, a single first- and second-level class summarized large numbers of retrievals. Summarization by classification numbers could be helpful to users might want to partition their retrievals to these single, high-posted first- and second-level classes or to first- and second-level classes that treated their topics of interest from a less-than-common perspective.

Generalizations about classification terminology:

- Generally, DDC terminology would be understandable to end users.
- LCC terminology was sometimes not suitable for conveying the subjects of classification numbers which were assigned to many titles bearing the same subject headings. Considerable editorial work would be necessary to improve the wording of captions at all levels of the Library of Congress Classification.

Classification-based summarization could be used in lieu of the exact approach. Systems could use broad ranges of classification numbers to consolidate retrievals and captions from the classification schedules to summarize the subjects of consolidated retrievals. When retrievals summarized by second-level captions are unmanageable (for example, greater than 50 titles), we recommend the use of subject headings to summarize them by subject.

On occasion, it is likely that all the typical methods of summarization — classification, subject headings, year of publication, language — would be exhausted and numbers of retrievals would still be unmanageable. To facilitate browsing in such searches, catalogs should feature display techniques that allow users to browse retrievals as quickly as possible. For example, the capabilities of "balloon help" from the Apple Macintosh's graphical user interface and of "show URL's" from Mosaic's graphical user interface could be extended to online catalog interfaces. When end users dragged a mouse over particular locations on the screen, detailed explanations of the underlying text would be given on the screen. Imagine dragging a mouse over a brief-titles list bearing truncated author names, call numbers, and titles. When you dragged the mouse over the titles, the system would display the full titles in a balloon overlay. In this way, end users could drag the mouse down a column of a dozen truncated titles, call numbers, author names, etc., and read the information that interests them in balloon overlays in a matter of seconds.

7.6Using Classification to Increase Very Low Retrievals

On occasion, keyword matches of end-user queries or truncated query words resulted in very low retrievals, i.e., less than ten titles. We culled queries from lists of low-posted, keyword-in-record searches to explore how classification could increase very low retrievals. Although we submitted the same queries to Michigan's MIRLYN online catalog and Duke University's online catalog, some queries retrieved adequate numbers of keyword-in-record retrievals in one of the two catalogs. In addition, the two databases contained different records, and, thus, resulted in different retrievals. We eliminated queries from the analysis that produced an adequate number of retrievals in keyword-in-record searches; consequently, we do not have results for the same queries as we had in sections 7.3 and 7.4 of this report.

Table 7.45 summarizes four searches in which browsing titles in the same Dewey Decimal Classification area as retrieved, useful ones led to the retrieval of additional useful titles. The left-hand column gives the original query followed by terms used in the keyword-in-record search; the plus sign (+) indicates where explicit truncation was applied.

Table 7.45. Using DDC Numbers to Increase Retrievals

Query/Search statement	Explanation
arts fund raising/ arts fund raising	3 of 5 titles retrieved in a keyword-in-record search are potentially useful: *ArtsMoney: raising it, saving it, and earning it, Successful fundraising for arts and cultural organizations,* and *Cash in!: funding and promoting the arts.* All 3 titles bear the class number "700.681." Searches of this class number result in two more titles, viz. *National guide to funding in arts and culture* and *Nonprofit enterprise in the arts.*

gays in the military/ gay+ and military	5 of 23 titles retrieved in a keyword-in-record search are potentially useful: *Conduct unbecoming: lesbians and gays in the U.S. military: Vietnam to the Persian Gulf*, *Gays — in or out?: the U.S. military & homosexuals: a source book*, *Barrack buddies and soldier lovers: dialogues with gay young men in the U.S. military*, *Gays and the military: Joseph Steffan versus the United States*, and *My country, my right to serve: experiences of gay men and women in the military*. All five are assigned the number 355.008664 or 355.0086642. Searches of these class numbers result in only one more title, viz. *Torn allegiances: the story of a gay cadet*.
banned books/ book+ and ban?	2 of 19 titles retrieved in a keyword-in-record search are potentially useful: *Banned books* and *Censored: books and their right to live*. Both books are assigned the number 098.1. Searches of this class number result in three more titles, viz. *Censorship: 500 years of conflict*, *Book burning*, and *Banned books, 387 B.C. to 1978 A.D.*
deaf and culture/ deaf and culture	1 of 2 titles retrieved in a keyword-in-record search are potentially useful: *Deaf in America: voices from a culture.* Searches of its class number (362.420973) result in two promising titles, viz. *Dancing without music: deafness in America* and *At home among strangers.*

Table 7.46 summarizes five searches in which browsing titles in the same Library of Congress Classification area as retrieved, useful ones led to the retrieval of additional useful titles. The left-hand column gives the original query followed by terms used in the keyword-in-record search; the plus sign (+) indicates where explicit truncation was applied.

Table 7.46. Using LCC Numbers to Increase Retrievals

Query/Search statement	**Explanation**
arts fund raising/ arts fund raising	3 of 5 titles retrieved in a keyword-in-record search are potentially useful: *ArtsMoney: raising it, saving it, and earning it*, *Successful fundraising for arts and cultural organizations*, and *Cash in!: funding and promoting the arts.* All 3 titles bear the class number "NX765." Searches of this class number result in three more titles, viz. *The emerging arts: management, survival, and growth*, *Arts administration: how to set up and run successful nonprofit arts organizations*, and *Arts administration and management: a guide for arts administrators and their staffs.*

gays in the military/ gay+ and military	2 of 16 titles retrieved in a keyword-in-record search are potentially useful: *My country, my right to serve: experiences of gay men and women in the military* with the class number UB418 .G38 and *Fighting back: lesbian and gay draft, military, and veterans issues* with the class number UB343. Searches of the former class number result in three more titles, viz. *Armed forces informer*, *Gays in uniform: the Pentagon's secret reports*, and *Exclusion: homosexuals and the right to serve.*
women and dieting/ women and dieting	The only title retrieved in a keyword-in-record search is potentially useful: *Beyond dieting: psychoeducational interventions for chronically obese women.* Searches of its class number RC552 .O25 result in four promising titles *Fat oppression and psychotherapy: a feminist perspective, Such a pretty face: being fat in America, Fat is a feminist issue: the anti-diet guide to permanent weight loss*, and *Weight, sex, and marriage: a delicate balance.*
introduction to polymer science/ introduction and polymer and science	3 of 7 titles retrieved in a keyword-in-record search are potentially useful, viz. *Introduction to physical polymer science*, *Introduction to polymer science and technology: an SPE textbook*, and *Introduction to polymer science.* (Only the former title was written in the 1990s.) Searches of these 3 books' class number, QD381, result in 4 promising, recent titles from a list of over 100 titles, viz. *Contemporary polymer chemistry* (1990), *Introduction to synthetic polymers* (1994), *Polymer chemistry: an introduction* (1992), and *Introduction to polymers* (1991).
lie group representation/ lie and group and representation	2 of 3 titles retrieved in a keyword-in-record search are potentially useful, viz. *Weil's representation and the spectrum of the metaplectic group* and *Spherical functions on a semi-simple lie group*. Searches of their class number, QA387, result in several promising titles from a list of over 150 titles, viz. *Lie groups and Lie algebras, Representations of nilpotent Lie groups and their applications, On the structure and complex representation theory of finite groups of Lie type,* and *Representations of finite groups of Lie type.*

The nine searches summarized in Tables 7.45 and 7.46 describe seemingly successful searches in which class numbers assigned to titles retrieved in keyword-in-record searches led to the retrieval of additional, potentially useful titles. These two tables fail to show the following problems with this strategy for finding additional useful titles:

- The many searches we conducted in which class numbers did not lead to potentially useful titles.
- The many searches we conducted in which keyword-in-record retrievals led to a handful of useful titles that were all assigned different class numbers.

- The many class number searches we conducted which led to so many titles with the same class number that browsing title lists to find potentially useful ones was too tedious and time-consuming.
- The many class number searches we conducted in which titles in the same classification area were more general than the topics users sought.

For the most part, searching classification numbers extracted from titles retrieved in low-posted keyword-in-record searches was a hit-or-miss proposition. This strategy could result in the retrieval of additional, potentially useful titles. It could fail to yield useful retrievals or result in retrievals that were more general than the topics users had in mind. It could also result in requiring patience and perseverance of end users who must browse many retrievals to find promising ones. Hildreth (1992) and Walker and De Vere (1990, 67) experimented with browsing using classified lists of titles and reported comparable results. When such browsing yields many titles, browsing techniques recommended for browsing large retrievals would also facilitate browsing classified lists of titles (section 7.5).

The place of classification-based browsing to overcome the problem of “too few retrievals” in the redesigned search trees would be as a search of “last resort.” We would tack it onto the one-word (figure 5.3B) and multi-word trees (figure 5.4C) following the alphabetical approach.

7.7Chapter Summary

The problem of too many retrievals plagued exact searches. Although the exact approach was intended to summarize large retrievals, it would be limited in usefulness because only three broad categories could be constructed from the three different types of subject subdivisions coded in subject heading fields of bibliographic records. Considerable editorial and developmental work efforts would be needed to fully implement the exact approach in online bibliographic systems.

Library classifications hold considerable promise for summarizing the results of high-posted searches in terms of their subject matter. Systems could use broad ranges of classification numbers to consolidate retrievals and captions from the classification schedules to summarize the subjects of consolidated retrievals. Chapter 7 explores the use of library classifications for consolidating and summarizing large numbers of retrievals.

We chose five moderately high-posted subjects entered by SULIRS, ORION, LS/2000, or MIRLYN users: (1) acid rain, (2) costa rica, (3) greek sculpture, (4) pornography, and (5) racism. All five queries were exact matches of controlled vocabulary terms. We searched these subjects in two sizable online catalogs: (1) Duke University’s online catalog using its subject heading search (s=) that listed retrieved titles for subdivided and unsubdivided forms of the matched subject heading, and (2) The University of Michigan’s MIRLYN online catalog using its subject heading search (s=) that listed retrieved titles for subdivided and unsubdivided forms of the matched subject heading. Retrievals in the Duke and Michigan catalogs were classified in the Dewey Decimal and Library of Congress Classifications, respectively. We consolidated retrievals in classification order and summarized large numbers of retrievals using captions from classification summaries and outlines (sections 7.3 and 7.4).

Numbers of titles retrieved for these five subjects in Duke's online catalog ranged from a low of 111 titles ("pornography") to a high of 497 titles ("costa rica"); numbers of titles retrieved for these five subjects in Michigan's online catalog ranged from a low of 141 titles ("acid rain") to a high of 476 titles ("racism"). The average number of retrievals per unique, retrieved DDC classification number ranged from a low of 2.1 titles ("racism") to a high of 5.2 titles ("greek sculpture") and per LCC classification number ranged from a low of 2.1 titles ("racism") to a high of 4.8 titles ("greek sculpture"). The lowest percentages of titles retrieved for the most common DDC number and LCC number for a particular subject were 11.8% for the synthesized DDC number "305.800973" under "racism" and 11.3% for "HC143" and "HT1521" under "costa rica" and "racism," respectively. The highest percentages of titles retrieved for the most common DDC number and LCC number for a particular subject were 42.5% for the DDC schedules number "733.3" under "greek sculpture" and 51.8% for the LCC number "TD196.A25" under "acid rain." Although our analysis of subject heading and classification number retrievals was far from comprehensive, the analysis resulted in several generalizations.

Although our analysis of subject heading and classification number retrievals was far from comprehensive, the analysis resulted in generalizations about retrievals, library classification classes, and classification terminology (section 7.5). The most important generalizations were:

- Retrievals were distributed into many unique classification numbers across many broad disciplines.
- Retrievals were likely to result in one common classification number per subject; however, the percentage of retrievals for such numbers could vary considerably, i.e., between ten and fifty percent of the titles assigned the subject heading.
- Generally, DDC terminology would be understandable to end users.
- LCC terminology was sometimes not suitable for conveying the subjects of classification numbers which were assigned to many titles bearing the same subject headings. Considerable editorial work would be necessary to improve the wording of captions at all levels of the Library of Congress Classification.

Classification-based summarization could be used in lieu of the exact approach. Systems could use broad ranges of classification numbers to consolidate retrievals and captions from the classification schedules to summarize the subjects of consolidated retrievals. When retrievals summarized by second-level captions are unmanageable (for example, greater than 50 titles), we recommend the use of subject headings to summarize them by subject.

Our exploration of the use of classification to enhance the results of searches that produced too few retrievals concluded that searching classification numbers extracted from titles retrieved in low-posted keyword-in-record searches was a hit-or-miss proposition (section 7.6). This strategy could result in the retrieval of additional, potentially useful titles. It could fail to yield useful retrievals or result in retrievals that were more general than the topics users had in mind. It could also result in requiring patience and perseverance of end users who must browse many retrievals to find promising ones.

The place of classification-based browsing to overcome the problem of “too few retrievals” in the redesigned search trees would be as a search of “last resort.” We would tack it onto the one-word (figure 5.3B) and multi-word trees (figure 5.4C) following the alphabetical approach.

The problem of too many retrievals also called for improved methods of browsing titles. On occasion, the use of classification to increase retrievals in searches plagued by too few retrievals would have to enlist improved methods of browsing titles. We recommended capabilities similar to the Apple Macintosh’s “balloon help” and Mosaic’s “show URL’s” in which dragging a mouse over truncated information would produce balloon overlays that showed the desired information in its entirety.

7.8References

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

Hildreth, Charles R. 1992. *An evaluation of structured navigation for subject searching in online catalogues.* Ph. D. dissertation, The City University, London.

Larson, Ray R. 1989. “Managing information overload in online catalog subject searching.” In *ASIS 89: Proceedings of the 52nd ASIS annual meeting*, 129–35. Medford, NJ: Learned Information.

Lynch, Clifford. 1990. “Large database and multiple database problems in online catalogs.” In *Annual review of OCLC Research, July 1989–June 1990*, 51–5. Dublin, OH: OCLC.

Prabha, Chandra. 1990. “Managing large retrievals: a problem for the 1990s? In *OPACs and beyond; Proceedings of a joint meeting of the British Library, DBMIST, and OCLC*,, 33–8. Dublin, OH: OCLC.

Wajenberg, Arnold. 1983. “MARC coding of DDC for subject retrieval.” *Information Technology and Libraries* 2 (September): 246–51.

Walker, Stephen, and Rachel De Vere. 1990. *Improving subject retrieval in online catalogues. 2. Relevance feedback and query expansion.* London: British Library. British Library Research Paper 72.

Wiberley, Stephen E., Jr., Robert Allen Daugherty, and James A. Danowski. 1993. *User persistence in displaying LUIS postings: a report to the Council on Library Resources.* Chicago: University of Illinois at Chicago.

8 Highlights of Project Activities and Findings

8.1Project Overview

Search trees are a set of paths with branches or choices that enable systems to carry out the most sensible search approach at each stage of the search. A new design for subject access to online catalogs enlisted search trees to identify the characteristics of end-user queries for subjects, control system responses, and determine appropriate subject searching approaches in response to the subject queries users entered into online catalogs.

The search trees that were the focus of this research project were developed from an empirical study of the subject queries users enter into online catalogs (Drabenstott and Vizine-Goetz 1990; 1994; Vizine-Goetz and Drabenstott 1991). Search trees were limited to subject searching approaches implemented in *operational* online catalogs, i.e., exact, alphabetical, and various keyword approaches. These approaches, however, failed to produce retrievals for some queries.

8.2Research Questions and Methods

The objective of this research project was to enhance the existing configuration of search trees with new subject searching approaches that were not available in operational online catalogs to provide useful information for the most difficult user queries. The study answered five research questions:

1. To what extent did user queries fail to produce retrievals through the subject searching approaches in the existing search-tree configuration?
2. What subject searching approaches would provide useful retrievals for these failed queries?
3. To what extent did user queries match controlled vocabulary terms that were not posted in the online catalog searched?
4. What subject searching approaches would provide useful retrievals for user queries that match unposted controlled vocabulary terms?

5. What enhancements were needed to the existing search-tree configuration to improve the quality and responsiveness of online catalogs to the user queries selected for study in this project?

In a previous research project sponsored by the Council on Library Resources, the Michigan project team performed a manual analysis of over fifteen hundred subject queries from the SULIRS, ORION, and LS/2000 online catalogs at Syracuse University, University of California, Los Angeles (UCLA), and University of Kentucky, respectively. This analysis resulted in the development of the original configuration of search trees (Drabenstott and Vizine-Goetz 1990; 1994). In this research project, the project team analyzed the same set of queries and analyzed over four hundred additional queries from the transaction logs of MIRLYN, the online catalog of the University of Michigan. The objective of the latter analysis was to answer the five research questions listed above about enhancing the existing configuration of search trees.

The Michigan project team selected the initial queries users entered in subject searches from the four libraries' transaction logs (sections 1.4 and 2.2). The team categorized queries by the type(s) of elements present in them: (a) topical subjects, (b) corporate names, (c) geographic names, (d) personal names, and (e) combinations of two or more elements (a-d).

The team developed subcategories of selected queries corresponding to the search-tree subject searching approaches that would provide useful retrievals. Staff also scrutinized subcategories of subject queries to determine whether they had certain characteristics that online systems could recognize without the help of human intermediaries. Search trees could perform the same operations on their own and pass queries with certain characteristics to newly-defined, subject searching approaches for which they were suited.

8.3Subject Searches in Systems Governed by Search Trees

Table 8.1 summarizes subject searches featured on search trees for subject queries generally and for personal-name queries according to the number of queries in this study that possessed the particular characteristics for the particular type of search.

Table 8.1. Matches of User Queries and Subject Searches

Subject search	Number	Percentage
Search trees for subjects generally		
Exact approach	832	43.4
Alphabetical	155	8.1
Keyword-in-main-heading	24	1.3
Keyword-in-subdivided-heading	74	3.9
Title-keyword	248	12.9

Keyword in subject heading fields	11	0.6
Keyword-in-record	31	1.6
No match	90	4.7
Subtotal (subjects generally)	1465	76.5
Search trees for personal names		
Keyword-in-subdivided-heading	2	0.1
Keyword-in-record	5	0.3
Alphabetical	223	11.6
Wrong tree	21	1.1
Subtotal (personal names)	251	13.1
Discarded queries	200	10.4
Total	1,916	100.0

Search trees for subjects generally handled a little over three-quarters of end-user queries. The largest percentage of searches (43.4%) would be given to the exact approach. Coming in a distant second place and third place were title-keyword searches (12.9%) and the alphabetical approach (8.1%), respectively. In an online bibliographic system governed by search trees, a small percentage of end-user queries would not be satisfied by the existing configuration of search trees. These queries were given in the “no match” category and accounted for only 4.7% of end-user queries.

Search trees for personal names handled 13.1% of end-user queries. A very small percentage (0.4%) of personal-name queries would retrieve titles through keyword searches. Most (11.6%) would be given to the alphabetical approach.

Discarded queries accounted for a little over 10% of end-user queries.

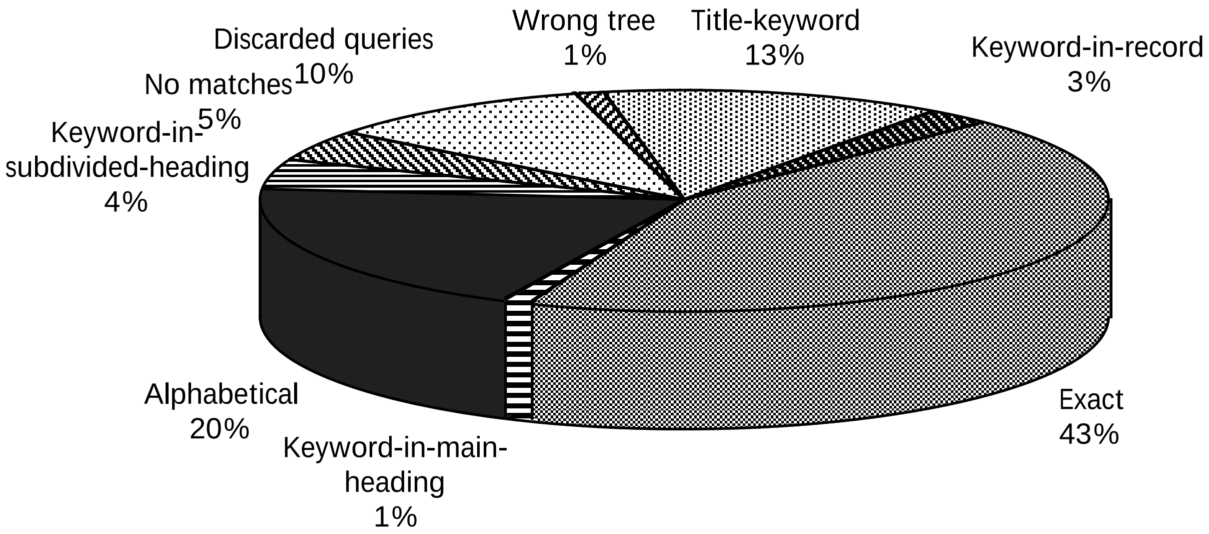


Figure 8.1. Matches of user queries and subject searches

Figure 8.1 consolidates queries and search-tree searches (and rounds percentages). Since exact searches were only featured in the search tree for subjects generally, the large percentage of user queries given to this search did not change — 43.4%. The percentage of alphabetical searches went up to 19.7% because both search trees featured this search. Over two-thirds of user queries met the criteria for controlled vocabulary searches (exact, alphabetical, keyword-in-heading, keyword-in-subdivided). No matches and queries handled by the wrong search tree accounted for less than 6% of user queries. Discarded queries were about 10% of user queries.

8.4Enhancements to Search Trees

Of the 1,919 queries for subjects generally, 90 (4.7%) failed to meet the criteria for exact, alphabetical, keyword-in-heading, and keyword matches (section 5.2). The majority (78.9%) of non-matches were queries for topical subjects (Table 5.1). Non-matches figured into lengthy searches that averaged four access points per search. They averaged over three words per query and failed to retrieve titles through the existing search-tree configuration.

We used an enhanced search-tree configuration from a related study to determine its effectiveness responding with useful retrievals to queries for which the original search-tree configuration was unable to produce retrievals (Drabenstott and Weller 1994, chapter 13).

The initial search tree remained basically unchanged from the original initial search tree (figure 5.2). The only change was to return users whose searches failed to produce retrievals to the question about personal names. Redesigned search trees for one-word and multi-word queries contained several changes from the original search trees (figures 5.3A–5.3B and 5.4A–5.4C): (1) checking go/see lists, (2) redirecting misspelled queries to the question on personal names, (3) invoking the keyword-in-record search (added to one-word tree), (4) invoking relevance feedback following keyword searches (i.e., in title-keyword, keyword in subject heading fields, and keyword-in-record searches), (5) invoking stemming (multi-word search tree only), (6) invoking the best-match approach (multi-word search tree only), and (7) invoking the alphabetical approach as the search type “of last resort.”

With these changes to the original search-tree configuration in mind, we divided non-matching queries into three groups according to the number of significant words in them: (1) one-word, (2) two-word, and (3) more than two-word queries. We then determined how non-matching queries would fare in systems that were governed by the enhanced search-tree configuration (section 5.4).

One-word and several two-word non-matching queries showed promise in terms of retrieving useful information through alphabetical searches. Truncation, spelling correction, and the availability of a go/see list with irregular plurals combined to produce retrievals for several two-word queries. Truncation, spelling correction, the availability of a go/see list with proper adjectives, and best-match approach combined to produce satisfactory retrievals for several queries that exceeded two words. Searching subject- or form-specific databases (e.g., Medline, Psychological Abstracts, National Newspaper Index) for several non-matching queries composed of two or more words was also a successful strategy for producing retrievals. No amount of manipulation produced useful results for a small number of queries.

8.5Handling Too Many Retrievals

To investigate the potential of library classifications for summarizing the results of high-posted searches in terms of their subject matter, we chose five moderately high-posted subjects entered by SULIRS, ORION, LS/2000, or MIRLYN users: (1) acid rain, (2) costa rica, (3) greek sculpture, (4) pornography, and (5) racism. All five queries were exact matches of controlled vocabulary terms. We searched these subjects in the online catalogs of Duke University and The University of Michigan because they were large databases in which retrievals were classified in the Dewey Decimal and Library of Congress Classifications, respectively. We consolidated retrievals in classification order and summarized large numbers of retrievals using first- and second-level captions from classification summaries and outlines (sections 7.3 and 7.4).

Although our analysis of subject heading and classification number retrievals was far from comprehensive, the analysis resulted in generalizations about retrievals, library classification classes, and classification terminology (section 7.5). The most important generalizations were:

- Retrievals were distributed into many unique classification numbers across many broad disciplines.
- Retrievals were likely to result in one common classification number per subject; however, the percentage of retrievals for such numbers could vary considerably, i.e., between ten and fifty percent of the titles assigned the subject heading.
- Generally, DDC terminology would be understandable to end users.
- LCC terminology was sometimes not suitable for conveying the subjects of classification numbers which were assigned to many titles bearing the same subject headings. Considerable editorial work would be necessary to improve the wording of captions at all levels of the Library of Congress Classification.

Classification-based summarization could be used in lieu of the exact approach. Systems could use broad ranges of classification numbers to consolidate retrievals and captions from the classification schedules to summarize the subjects of consolidated retrievals. When retrievals summarized by second-level captions were unmanageable (for example, greater than 50 titles), we recommend the use of subject headings to summarize them by subject.

8.6Handling Too Few Retrievals

Our exploration of the use of classification to enhance the results of searches that produced too few retrievals concluded that searching classification numbers extracted from titles retrieved in low-posted keyword-in-record searches was a hit-or-miss proposition (section 7.6). This strategy could result in the retrieval of additional, potentially useful titles. It could fail to yield useful retrievals or result in retrievals that were more general than the topics users had in mind. It could also result in requiring patience and perseverance of end users who must browse many retrievals to find promising ones.

The place of classification-based browsing to overcome the problem of “too few retrievals” in the redesigned search trees would be as a search of “last resort.” We recommend tacking it onto the one-word (figure 5.3B) and multi-word trees (figure 5.4C) following the alphabetical approach.

8.7 Improving Long Lists of Retrieved Titles

The problem of too many retrievals called for improving methods of browsing titles. On occasion, the use of classification to increase retrievals in searches plagued by too few retrievals would have to enlist improved methods of browsing titles. We recommended capabilities similar to the Apple Macintosh's "balloon help" and Mosaic's "show URL's" in which dragging a mouse over truncated information would produce balloon overlays that showed the desired information in its entirety.

8.8 References

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

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

Drabenstott, Karen M., and Marjorie S. Weller. 1994. *Testing a new design for subject searching to online catalogs.* Ann Arbor, MI: School of Information and Library Studies, University of Michigan.

Vizine-Goetz, Diane, and Karen M. Drabenstott. 1991. "Computer and manual analysis of subject terms entered by online catalog users." In *Proceedings of the 54th ASIS annual meeting*, edited by Jose-Marie Griffiths, 156–61. Medford, NJ: Learned Information.

Appendix A.
DDC Retrievals for “acid rain”

Class numbers retrieved in subject heading searches of Duke University’s online catalog for “acid rain” are followed by the number of titles retrieved bearing the listed class number.

016.3092	2
016.3637386	6
016.628168	1
328.769	1
341.7623	4
341.7625	2
344.046342	4
354.710082	1
363.1683	1
363.73846	2
363.7386	35
363.7386097	4
363.73862	2
363.73865	4
363.738656	3
363.739	1
363.7392	20
363.7394	12
551.5781	3
570.8	1
574.5222	1
574.5263	1
574.52632	1
581.52642	1
628.16	1
628.168	2
628.5	4
628.53	2
628.532	4
631.42	2
634.9	1
634.906	1
639.20971	1
639.3	1
660.8	2
741.8946	1
971.064	1

Appendix B.
DDC Retrievals for “costa rica”

Class numbers retrieved in subject heading searches of Duke University’s online catalog for “costa rica” are followed by the number of titles retrieved bearing the listed class number.

015.7286	3
016.3058	1
016.9127286	3
016.97286	2
261.8097286	2
282.092	3
282.7285	1
282.7286	6
301.36	1
303.484	1
303.6097286	1
304	2
304.6097286	5
304.64	1
305.5097286	3
305.52	2
305.5633	1
305.896	1
306	1
306.097286	3
306.4209728	1
308.2	1
309.17286	1
312.097286	13
317.286	16
320.12	1
320.5315	2
320.5409728	1
320.7286	1
320.97286	7
320.98	2
321.8	1
321.8097286	1
323.042	1
324.27286	5
324.98	2
327.08	2
327.7284	1
327.7285	1
327.7286	13
327.7286072	2
327.7307286	1
327.861	2
328.7286	1
330.9	3
330.97286	51
330.98	4
330.9862	2
331.097286	1
331.098	2
331.1098	1
331.125	1
331.5245	1
331.833	1
332.1097286	1
332.11	2
332.152	2
333.72	1
333.7309728	1
333.8809	1
336.3409728	1
336.47286	1
336.7286	1
338.097286	10
338.47	1
338.91	1
338.9173	1
338.97285	1
338.97286	8
339.2097286	1
341.2	2

341.42	1
341.442	3
341.6	1
342.7286	4
342.7286023	2
342.7286029	2
342.7286066	1
352.0072	1
352.087286	2
354.728073	1
354.7286	3
354.7286008	1
354.728601	1
354.728603	1
354.7286035	2
354.7286063	2
354.7286092	1
355.0097286	4
355.0213	1
355.3097286	2
361.6109728	2
363.5097286	1
363.7097286	1
370.97286	1
378.1981	1
381.41373	1
382.097286	3
394.097286	1
572.08	1
581.97286	1
614.097286	1
630.97286	2
709.7286	1
736.24	1
860.997286	1
861	2
912.7286	2
912.97286	2
913.7	1
913.7286	1
917.286	9
917.2860453	1
920.07286	1
923.27286	1
923.57286	1
923.946	1
927.8053	1
929.5097286	1
962.8605092	1
970.1	2
972	5
972.8	1
972.804	1
972.85	1
972.85044	1
972.86	78
972.86003	4
972.8600497	1
972.8601	1
972.8602	1
972.8603	6
972.8604	12
972.86044	2
972.8605	16
972.8605092	5
972.86051	12
972.86052	16
972.86053	2
972.86092	3
980	1
986.14	1

Appendix C.
DDC Retrievals for “greek sculpture”

Class numbers retrieved in subject heading searches of Duke University’s online catalog for “greek sculpture” are followed by the number of titles retrieved bearing the listed class number.

063.3	1
063.6	1
069.3	1
704	1
708.3	1
708.315	4
708.891	1
709.3	3
709.37	1
709.38	5
709.45	1
709.489	1
721.273	1
722.8	5
726.80938	1
730	1
730.0938	1
730.92	2
731.76	4
732.916	1
733	71
733.3	121
733.30937	1
733.309385	1
733.3093915	1
733.309392	1
733.3154	1
733.31732	2
733.38	1
733.388	1
733.5	2
735	1
737	1
806	1
880.4	1
880.9	1
906.46	1
913.08	2
913.3	8
913.32	1
913.37	4
913.38	9
913.385	1
913.387	1
913.388	2
913.3915	1
913.8	1
913.85	1
914.951	1
937.7	1
938	2
939.14	1
939.8	1
949.96	1
956.3	1

Appendix D.
DDC Retrievals for “pornography”

Class numbers retrieved in subject heading searches of Duke University’s online catalog for “pornography” are followed by the number of titles retrieved bearing the listed class number.

016.36283	1
016.36347	6
025.273067	2
028	1
108	1
132.75	1
176.8	1
204	1
241.667	1
248.246	1
261.8	1
286.1092	1
301.162	1
301.21	1
302.23	1
305.4	1
305.40952	1
305.42	2
305.420973	1
306.7	1
306.7094	1
306.7420971	1
306.77	3
340.8	1
344.0547	1
344.730547	3
363.310944	2
363.440971	1
363.47	44
363.4709	2
363.470952	1
363.470971	3
363.4709711	1
363.470973	5
363.4709753	1
364.174	3
467.09	1
704.9428	2
791.4363538	1
809.933538	1
840.9	1
848.91409	1
947	4

Appendix E.
DDC Retrievals for “racism”

Class numbers retrieved in subject heading searches of Duke University’s online catalog for “racism” are followed by the number of titles retrieved bearing the listed class number.

016.3058	1
016.305896	1
016.97304	1
155.7	1
155.82	2
157.3	2
191	1
209.73	1
230	1
261	1
261.8	1
261.8348	5
261.834896	1
266.22	1
277.3081	2
280	1
284.173	1
299	1
301.1543	1
301.1832	1
301.45	4
301.451	6
303.385	2
303.387	1
304.2	1
304.5	3
304.8094542	1
304.871	2
305	2
305.0973	1
305.3	2
305.4	3
305.42	2
305.420973	2
305.48	1
305.488034	1
305.48896	1
305.5009	1
305.50973	1
305.5209728	1
305.8	40
305.8009	3
305.800904	1
305.80092	2
305.80094	7
305.800941	11
305.800942	1
305.8009424	1
305.8009427	1
305.800943	3
305.800944	2
305.8009494	1
305.800954	1
305.800968	3
305.80097	1
305.800971	8
305.800973	41
305.80098	1
305.8009969	1
305.8034073	1
305.8094	1
305.80943	1
305.80945	1
305.86	1
305.890711	2
305.891411	1
305.894811	1
305.896	2
305.896037	1
305.896041	2

305.8960492	1
305.896073	25
305.896081	1
305.96073	1
305.986	1
306.362	1
306.36209	1
306.483	1
320.019	1
320.50944	1
320.52094	1
320.533	1
320.5330943	2
320.56	5
320.56092	3
320.560941	1
320.5609931	1
321.04094	1
322.420973	2
323.1	3
323.108	1
323.110941	1
323.1196073	1
323.142	2
323.168	1
323.171	1
324.0973	1
324.273409	1
325.3	1
325.43	1
325.73	1
327.11	1
330.122	1
330.968063	1
331.133	1
331.544	1
331.6097286	3
331.62561	1
331.6396073	1
342.0873	2
342.750873	1
345.73	1
346.73013	1
347.7302523	1
347.732634	1
355.08	1
361.610941	1
363.7008693	2
363.92	1
364	1
364.153	1
370.115	1
370.190941	1
370.19342	2
371.81	1
378.687	2
378.782	1
379.41	1
572.092	1
572.2	1
668	1
784.6	1
811.54	1
813.54	2
904	1
909.0496073	1
909.82	1
940.53	1
940.531	2
941.082	2
943.084092	1
943.605	1
944.004924	1
944.00493	1
944.0838	1
970.004	1
971.300496	1
972.004951	1
973.0496	1
973.0496073	3
973.9092	1
973.917	1
973.92092	1
975	1
975.04	1
975.304	1
975.8004	1
976.1781063	1
977.7	1
979.494053	1
980	1
981.00496	1
985.063	1

Appendix F.
LCC Retrievals for "acid rain"

Class numbers retrieved in subject heading searches of the University of Michigan's online catalog for "acid rain" are followed by the number of titles retrieved bearing the listed class number.

HC79.E5	2
HD9685.C2	4
JK5374.A32	1
K3584.6	1
K3593	2
QC851 .W93	1
QC861.2	1
QC924.5	1
QC981.8.C5	2
QH104.5.N58	1
QH541.5.F6	1
QH545.A17	20
QH545.A3	1
QK326	1
QL618.3	1
QP1	1
QR105.5.A191	1
S593.5.K461	2
SH37.A35	3
TA418.74	1
TD195.4	2
TD195.42	3
TD195.5	2
TD195.54.C2	3
TD196.A25	73
TD427.A27	1
TD883.7.E8	1
TD885.5.S85	2
Z5862.2.A26	4
Z5862.2.A35	2

Appendix G.
LCC Retrievals for “costa rica”

Class numbers retrieved in subject heading searches of the University of Michigan’s online catalog for “costa rica” are followed by the number of titles retrieved bearing the listed class number.

AN68	1
AS36	2
AS65.C84	5
AS182	1
BX1436.2	5
BX1442.2	1
BX3612.5.C8	1
BX4373	3
BX8762.A45	2
CC65	1
CT275.K28	1
DT791	1
E11	1
E183.8.C8	1
F1401	2
F1403	2
F1434	1
F1437	1
F1438	1
F1439	1
F1526.22.C8	1
F1526.27	1
F1527	1
F1529.B7	4
F1529.S35	1
F1541	3
F1541.6	2
F1542	2
F1542.7	6
F1543	14
F1543.3.A7	1
F1543.5	2
F1544	5
F1545	7
F1545.3.A7	1
F1546	34
F1546.3	1
F1546.9	1
F1547	9
F1547.5	16
F1548	19
F1548.2	3
F1548.23	2
F1549.B7	33
G5	2
G58	3
G104.U55	1
G4860	9
G4861	1
GA491	1
GA514	2
GN2	1
H1	1
H31	2
HA801	2
HA802	2
HA804	2
HA805	1
HB3537	3
HC143	39
HC147	1
HC389	2
HD1531.C8	1
HD4024	1
HD9199.C8	1
HE2835	1
HF3251	1
HG2736	1

HG8553.5	1
HJ17	1
HJ2473	1
HJ8525	1
HM276	1
HN125.2.E4	1
HN133	3
HQ567	2
HQ1473	1
HX123.8.F38	1
J177	1
JL1403	1
JL1442	1
JL1443	4
JL1444	1
JL1449.A15	1
JL1456	1
JL1458	3
JL1459.A45	1
JL1459.A53	2
JL1459.A55	4
JL1459.A558	1
JX1283.U9	1
JX4033.C8	1
L150	1
LA446	1
LA448.7	1
PQ7488	1
PZ9	1
Q11	1
QE1	1
QH108	1
QK217	3
RA454.C8	1
SB484 .C8	1
Z1437	1
Z1451	3
Z1453	1
Z5114	1
Z6027.C8	1
Z6954.C8	4

Appendix H.
LCC Retrievals for "greek sculpture"

Class numbers retrieved in subject heading searches of the University of Michigan's online catalog for "greek sculpture" are followed by the number of titles retrieved bearing the listed class number.

AS32	1
AS162	4
AS182	7
AS222	4
AS242	1
AS281	1
BL785	1
CC5	2
CC9	1
CC65	1
CJ351	1
D5	5
DD901	1
DE2	12
DE3	2
DF77	1
DF78	1
DF135	1
DF221	3
DF261	3
DF285	1
DF287	2
DF287.A2	2
DF287.A3	1
DF918	1
DS41	3
DS155	2
DS156	1
DT57	1
N13	48
N25	1
N610	1
N2490	1
N3690	3
N5320	9
N5325	1
N5330	1
N5350	1
N5605	10
N5605.A67	2
N5610	1
N5630	3
N5633	1
N7445	1
N7570	1
N7585	3
N7587	1
NA260	1
NA270	1
NA280	1
NA281	1
NA285.P4	2
NA5060	1
NB27	3
NB70	1
NB71	3
NB82	1
NB85	7
NB86	2
NB87	18
NB90	98
NB91	49
NB92	4
NB93	3
NB94	26
NB98	2
NB100	3
NB102	2
NB104	1

NB105.A4	1
NB110	1
NB115	1
NB130	1
NB133	5
NB133.5	1
NB140	5
NB160	1
NB163	9
NB164	8
NB270	1
NB1135	3
NB1150	1
NB1312	1
NB1370	8
NB1390	1
NB1800	1
NB1880	1
ND140	1
NK4645	1
NK7951	1
PA25	1
Q11	1
QA90	1

Appendix I.
LCC Retrievals for “pornography”

Class numbers retrieved in subject heading searches of the University of Michigan’s online catalog for “pornography” are followed by the number of titles retrieved bearing the listed class number.

BH39	1
BX6495.F3	1
HE 6424	1
HG6185	1
HM258	1
HN80.W3	1
HN400.M3	2
HQ18.G7	1
HQ21	3
HQ29	1
HQ72.C6	1
HQ72.U53	1
HQ144	1
HQ148	4
HQ458	1
HQ460	1
HQ461	1
HQ471	94
HQ472	10
HQ1154	2
HQ1206	2
HQ1233	1
HQ1403	1
HV6249	1
HV6250.4.W65	1
HV6705	1
HV6695	1
HV6727	5
HV9076.5	1
J87	1
J2004	4
JQ769.5.C6	1
JX681	1
K5293	1
KF1262	1
KF4770.A75	4
KF9444	11
NX650.E7	1
PK2190	1
PN56.E7	1
PN56.S5	1
PN1995.9.S45	1
PR6023	1
PS3513.O527	2
PS3521	1
RC607.A26	2
Z657	1
Z659	9
Z688	2
Z7164.P845	3
Z7164.S42	1

Appendix J.
LCC Retrievals for “racism”

Class numbers retrieved in subject heading searches of the University of Michigan’s online catalog for “racism” are followed by the number of titles retrieved bearing the listed class number.

AS36	1
BF432.A1	3
BF432.N5	1
BF575.P9	1
BF723.R3	1
BF731	1
BM535	1
BR563.N4	3
BT82.7	1
BT734.2	3
BX1795.R121	1
CB195	1
CB235	1
CT275.N85	1
D767.9	1
D810.N4	2
D1056	4
D1056.2.B55	1
DA125	31
DA570	2
DA592	1
DA670.W495	1
DA690.L8	1
DA690.W86	1
DB2	1
DC33	1
DC34	8
DC34.5.N67	1
DC131	1
DC419.L58	1
DC419.L595	1
DD61.8	1
DD74	2
DD231	1
DD256.5	2
DD258.5	1
DD290.29	1
DJ92.S8	2
DP102	1
DP104	1
DR1524.S47	1
DS17	1
DS135.R7	1
DS145	1
DS146.U5	1
DS475	1
DS489.84	1
DS832.7.K6	1
DT1756	1
DU120	2
DU624.6	1
E29.A1	1
E29.N4	1
E175.85	1
E179.5	3
E184.A1	19
E184.C24	1
E184.7	1
E185	3
E185.2	4
E185.5	1
E185.61	12
E185.615	22
E185.625	3
E185.8	2
E185.86	5
E185.92	1
E185.93.G4	1
E185.93.I64	1

E185.97.C62	1
E185.97.D82	1
E185.97.J15	1
E185.97.K45	2
E185.97.S53	1
E415.7	1
E441	1
E446	1
E449	1
E667	1
E748.M985	1
E838	4
E838.L533	1
F213.R38	1
F215	1
F216.2.J341	1
F231.3.W55	1
F264.G8	1
F319.M6	1
F499.C6	1
F755.A1	1
F1035.A1	7
F1059.5.T689	1
F1059.7.A1	1
F1392.A1	1
F1789.N3	1
F1983.N4	1
F2082	1
F2151	1
F2471.A1	2
F2581	1
F2659.A1	1
F2659.N4	1
GF50	1
GN269	3
GN315	1
GN320	1
GN365.9	1
GN495.8	1
GN537	1
GV865	1
GV706.32	1
H1	1
H62	1
HC110.E5	2
HD6490.R2	1
HD8066	1
HD8039.B232	2
HD8081.A65	2
HD8398.B55	1
HM291	10
HN65	3
HN90.R3	1
HN90.S6	1
HN110.Z9V79	1
HN150.Z9	1
HN390	1
HN730.Z9	1
HQ75	7
HQ755.5.U5	1
HQ1031	1
HQ1090	1
HQ1154	3
HQ1237	1
HQ1421	1
HQ1426	2
HS2321	2
HS2330.K63	7
HS2330.N23	2
HS2330.O73	1
HS2330.O75	1
HT913	1
HT1501	2
HT1505	2
HT1507	1
HT1521	56
HT1523	4
HT1581	1
HV848.L552	1
HV4708	1
HV6485.G7	1
HV6534.A7	1
HV6568.A4	1
HV6791	1
HV8069	1
HX1	4
HX3	2
HX39.5	1
HX40	1
HX86	1
HX902	1
JA74.5	1
JA84.E9	2
JC311	1
JC481	1
JC599.C3	1
JC599.U5	4
JK2316	1
JN15	1
JN1129.N28	1
JN3007.F65	1

JN5981	1
JV7225	1
JV305	3
JV7225	1
JV7925	2
JV7928.R12	1
K3254.A56	1
K5274	2
KD4095	1
KF224.B73	1
KF3464	1
KF4757	1
KF4772	1
KF8205	2
KF8745.T48	1
KF9345	1
LA229	1
LA721.81	1
LB1028	1
LB3045	1
LB3045.6	1
LB3045.64	4
LB3047	1
LB3048.J3	1
LC93.G7	1
LC192.2	11
LC212	1
LC212.3.G7	5
LC212.53.G7	1
LC212.53.G73	1
LC214.23.B67	1
LC1099	1
LC1099.5.G7	1
LC1567	1
LC2701	1
LC2806.G7	3
LC3731	1
LD972.9	1
LD2222.35	1
LG471.U6	1
LH1	1
ML3556	1
NX650.R3	1
P120.R32	1
P120.S48	2
PE64.M38	1
PL659	1
PN56.R18	1
PN1009.Z6	1
PN2071.B58	2
PN3352.M5	1
PN5124.R28	1
PN5124.R29	2
PR6003.A544	1
PR6060.O38	1
RA644.L3	1
RA790.6	1
RC451.4.M58	1
RC451.5.A2	1
RC451.5.N4	1
RC455.4.E8	2
Z658.J3	1
Z682.2.U5	1
Z1361.E4	1
Z2247.M54	1
Z5817	1
Z7164.R12	5