DOCUMENT RESUME

ED 308 879 IR 052 876

| | |
|---|---|
| AUTHOR | Hildreth, Charles R. |
| TITLE | Intelligent Interfaces and Retrieval Methods for Subject Searching in Bibliographic Retrieval Systems. |
| INSTITUTION | Library of Congress, Washington, D.C. |
| REPORT NO | ISBN-0-8444-0626-0 |
| PUB DATE | 89 |
| NOTE | 127p.; Advances in Library Information Technology. Issue Number 2. |
| AVAILABLE FROM | Cataloging Distribution Service, Library of Congress, Washington, DC 20541. |
| PUB TYPE | Information Analyses (070) -- Viewpoints (120) |
| EDRS PRICE | MF01/PC06 Plus Postage. |
| DESCRIPTORS | *Artificial Intelligence; Computer Software; Information Retrieval; *Man Machine Systems; *Online Catalogs; *Online Searching; *Online Systems; Reference Services; *Search Strategies |
| IDENTIFIERS | Gateway Systems |

ABSTRACT

This study was designed to be a state-of-the-art survey and investigation of intelligent "front end" design approaches and software for improving subject access and subject searching in today's large online bibliographic retrieval systems and online public access catalogs (OPACS). The report begins with an illustrated overview of retrieval features and subject searching in current second-generation OPACs, which is followed by a discussion of the problems and shortcomings of conventional OPACs and online information retrieval systems, especially with regard to subject access and retrieval methods. A summary discussion of perspectives, models, and design contributions from the information retrieval research and experimentation community is then presented, and an illustrated review and analysis of intelligent interfaces and retrieval methods is provided. The discussion facilitates a refined grouping of the issues into four major areas: (1) ease of use, orientation, and presentation factors; (2) vocabulary control and correlation factors; (3) more effective system-guided or automatic query formulation and retrieval techniques; and (4) meaningful engagement of the searcher in relevance assessments, query modification/expansion, and provision of smart navigational, exploration facilities. A summary of recommended design principles and achievements in intelligent interface and retrieval system design concludes the report. A checklist of features of second-generation OPACs and a list of OPAC and retrieval systems and software investigated for the study are appended. (51 references) (MES)

# INTELLIGENT INTERFACES AND RETRIEVAL METHODS

## FOR SUBJECT SEARCHING IN BIBLIOGRAPHIC RETRIEVAL SYSTEMS

CATALOGING DISTRIBUTION SERVICE, LIBRARY OF CONGRESS

# INTELLIGENT INTERFACES AND RETRIEVAL METHODS

## FOR SUBJECT SEARCHING IN BIBLIOGRAPHIC RETRIEVAL SYSTEMS

Prepared by Charles R. Hildreth, READ Ltd.
for the Library of Congre‹

Advances in Library Information Technology
Issue Number 2

Cataloging Distribution Service, Library of Congress
Washington, D.C. 1989

Library of Congress Cataloging-in-Publication Data

Hildreth, Charles R., 1938-
Intelligent interfaces and retrieval methods for subject searching in bibliographic retrieval systems / prepared by Charles R. Hildreth for the Library of Congress.
p. cm. -- (Advances in library information technology, ISSN 0899-1227 ; issue no. 2)
Bibliography: p.
ISBN 0-8444-0626-0
1. Catalogs, On-line--Subject access. 2. On-line bibliographic searching. 3. Information retrieval. I. Library of Congress. II. Title. III. Series.
Z699.35.S92H55 1989 025.3'13--dc19 88-39045

ISBN 0-8444-0626-0

For sale by the Cataloging Distribution Service
Library of Congress, Washington, D.C. 20541 U.S.A.
(202) 707-6100

## CONTENTS

*"Looking is easy; finding presents the difficulty."*

-anon.

## 1. Introduction

The original aim of this study was to conduct a state-of-the-art survey and investigation of intelligent "front end" design approaches and software for improving subject access and subject searching in today's large online bibliographic retrieval systems, including online public access catalogs (OPACs). Special attention was to be directed to the problems and promise inherent in thesaurus or multithesauri-based subject searching in large online bibliographic databases and library catalogs. The major objective of the study was to discover, distill, or infer from these new "front end" design approaches worthy design principles which could be applied in the enhancement, redesign, or replacement of today's conventional online bibliographic retrieval systems. Thus, the major objective of the study gave license, if not direction, for enlarging the scope of the study beyond an investigation of online retrieval system front ends. In the course of the study this license was exercised because I agree with the view held by many researchers and information specialists that most front-end systems available are interim and narrow solutions to the subject access and retrieval problems associated with today's conventional information retrieval systems. (A note on usage: "information retrieval (IR) system" as used in this report is also known as "online database search system," "online reference retrieval system," "document retrieval system," or "bibliographic retrieval system." The first three are nearly synonymous; online bibliographic retrieval systems and OPACs are special purpose interactive information retrieval systems.) The meaning of "front end" with or without "intelligent" preceding it is ambiguous and its use in the OPAC and IR literature is not uniform and settled. Bates' suggestion for installing a "front-end system mind," a "dense semantic network," in OPACs is called a front end "because it is the part of the system the searcher encounters first."[1] Bates then explains that such a front end can play a useful role throughout the search process. A common understanding of the concept "front end" is that it refers to specially designed software that resides in front of, and outside of, the basic configuration boundaries of an actual stand-alone automated retrieval system. This view derives from the early front-end systems developed by innovative pioneers to ease access to the remote commercial online database search systems. Intended to make access and searching easier for untrained end users, these front ends were implemented on minicomputers or microcomputers which stood physically some distance from the commercial host systems, connected to the hosts only by telecommunication networks. Such front ends had no effect on the independent operation and structure of the host systems.

## INTRODUCTION

In time, the meaning of "front end" has shifted from the physical "frontal" dimension to a more metaphorical usage. Continued development of front-end software database access and search aids is largely explained by the wide use of microcomputers in online, interactive database searching in the 1980s. As Mischo explains, these aids were "designed to facilitate online searching by simplifying and automating the search process, with an eye to saving costs and improving search results."[2] However, further investigation reveals that this class of software -- referred to variously as "gateways," "transparent systems," "intermediary systems," and "expert interface systems" -- may be located in a variety of places in the overall retrieval system configuration. Mischo points out that front-end software can reside in four locations: "1) on a microcomputer; 2) on a vendor mainframe (BRS/AFTER DARK, KNOWLEDGE INDEX); 3) on remote dial-up computers (EasyNet); and on a local mainframe with direct access."[3]

"Front end" is becoming both richer in its meaning (functionally speaking) and more varied in its use. Generally it incorporates a variety of software-based approaches to making information retrieval systems more usable and effective for a variety of end users (i.e., users with an information need who search a retrieval system for themselves to resolve that need). For the most part these approaches involve practical solutions to "user-system interface/interaction" problems. Special-purpose user-system interface (customarily shortened to "interface") software is often employed to improve matters at that interface and, perhaps also, to make it a more "intelligent interface." Accordingly, to focus on solutions rather than where they may or may not physically reside, the preferred term in this report will be "intelligent interface" not "front end." Furthermore, in information retrieval research and development circles, no hard distinction is made between "intelligent interfaces" (or "front-ends") and intelligent information retrieval (IR) systems. The emphasis is on improved software, data resources, and database design and structure, not so much on the locus of the "intelligence" to be built into the overall information retrieval system or OPAC. In this community there seems to be a consensus on what constitutes, in general terms, an intelligent information retrieval system; the consensus is easily expressed in negative terms: today's conventional keyword-indexed, inverted file, Boolean logic search and retrieval systems like BRS, DIALOG, ELHILL, DATASTAR, BLAISE-ONLINE, LEXIS-NEXIS (and all second-generation OPACs) are powerful and efficient but are dumb, passive systems which require resourceful, active, intelligent human searchers to produce acceptable results.

Among information scientists there is a positive consensus on at least the directions we should pursue in designing more intelligent interfaces and intelligent IR/OPAC systems. In the simplest terms it involves transcending the known and proven limitations and shortcomings of today's conventional IR and OPAC systems. In positive terms it means improving both the functional performance and usability of operational IR and OPAC systems for a *variety* of searchers and search needs; it means pursuing design goals beyond merely meeting the requirements of trained search intermediaries and their search needs, needs currently supported by conventional retrieval systems. From the user interface to the design and structure of the database(s), today's inverted file, Boolean query systems, recognizably enhanced by word or field proximity search methods and menu-based command interfaces, are judged to be insufficient for the future and unqualified for the label "intelligent." At a deeper level, it is generally agreed that merely attaching a "friendly" user interface on a "front-end" processor to a dumb conventional retrieval system does not make the system more intelligent or effective. The twin goals pursued by most information scientists and IR system researchers are 1) improved and broader usability and 2) smarter functionality integrated into IR systems and OPACs.

Two additional "realities" must be explained. The problem of improving information retrieval systems by making them more intelligent and usable (and any subset of this problem, such as OPAC subject retrieval) is a *complex* problem. It is complex in both the number and the kind of subproblems that must be solved. There is no single, monolithic solution to the "problem" that can be materialized through unlimited software and/or hardware resources. Even the problem of OPAC subject access to MARC files has many dimensions. However, it is possible to identify a small set of experimental and design approaches which address different aspects of the problem of subject retrieval. Although investigated or developed in separate efforts and environments, they can be viewed as complementary thrusts toward more intelligent retrieval systems.

The second reality is that intelligent IR/OPAC system development is in its infancy. Most of the experimental and design activity has been confined to the academic or research environments. A small number of prototype or demonstration systems exist, and several notable "smarter" OPACs have been installed in "real user" environments. Using a narrow sense of "intelligent" as in "artificial intelligence" (meaning AI-based with a built-in 'knowledge base' and a rule-based inferential engine, for example, an 'expert system'), it is certainly true that intelligent IR projects are just beginning or are at early stages of development. However, for the scope of this study intelligent information retrieval begins where conventional online IR systems and second generation OPACs end. This casts the net a bit wider and further to gather significant findings and advances produced by IR research and recent OPAC developments. The range of advances in intelligent, more usable retrieval systems lies along a continuum from the most powerful conventional IR systems, to state-of-the-art prototypes, on to almost certain near-future design enhancements based on sound experimental research coupled with new and emerging technologies. As one cynic (or optimist) has put it: "while waiting for AI-based advances in intelligent systems, we can move from dumb systems to 'commonsense' systems."

IR and OPAC researchers find it unfortunate that most OPACs in operation have advanced only to become conventional IR systems, replacing or combining traditional catalog access methods with Boolean query features and word proximity search capabilities. Beneath a more palatable user interface, today's OPAC capabilities closely resemble the retrieval methods of the conventional systems like BRS, DIALOG, and ELHILL.

## INTRODUCTION

Fortunateiy, many IR researchers have taken an interest in OPACs and related "end-user" systems, seeing them as fertile ground for further experimentation and development. Their activities are moving us piecemeal but solidly into the next generation of OPACs and IR systems. These efforts can be grouped into three or four different but complementary approaches to making these systems more "intelligent" and usable. My point of departure must be kept in mind: with regard to functionality, usability, and performance, intelligent IR/OPAC systems begin where conventional systems end. Some researchers and writers have a more restricted, more specific view of intelligent systems. In addressing what makes a system intelligent, they may require that the system have a knowledge base and rule-governed inferential capabilities that can be used to make the appropriate connections between a request (typically in natural language) and a coilection of documents. If the knowledge base, rules, and logic are based on the knowledge and decision-making capabilities of real experts, the system is called an "expert" system. Building such expert OPAC systems is one approach to making OPACs more intelligent; like the other approaches to be described, it has exciting potential as well as inherent limitations.

Researchers have shown that the challenge of retrieving documents that will have the highest probability of being relevant to the user's information needs and/or interests is not one-dimensional. Add to these the challenge of making the system usable for direct use by a variety of users, both trained and untrained, experienced and inexperienced. Clearly, this is the situation facing us with improving subject retrieval in OPACs. It is not surprising that several approaches and different techniques are being applied to making OPAC and IR systems more intelligent. The emerging consensus is that progress lies in the direction of a combination of these approaches: employing features of the non-Boolean, probabilistic retrieval model, dynamic interaction with the user during the search process to gather evidence about relevance and preferences, plausible query term-document "closest" match inference methods (including natural language processing), and vastly improved presentations of data and assistance at the user interface.

More than twenty advanced OPACs and interactive bibliographic retrieval systems and software packages were investigated for this study (they are listed in Appendix 2). For a variety of reasons and circumstances they are not all equally intelligent. Some are complete stand-alone, operational retrieval systems, either university-based or vendor-supplied. Others are experimental prototypes or demonstration systems. A few cases of special purpose interface or retrieval experimental software were investigated. These systems and software projects operate in a large variety of computer environments. Some are installed and operating as the online catalog in large and small libraries. Others are maintained as special projects in educational research or experimental/test environments. Thus, comparisons among systems and software would not be meaningful. Progress at the cutting-edge of technology and new design can be very uneven. The systems were selected for investigation because they represent attempts to apply innovative, unconventional, "smart" design approaches and software techniques to the solution of known and confirmed interface and retrieval problems. What they have in common is a sound theoretical and empirical basis. To an unprecedented degree, they represent design and development responses to research findings, experimentally-acquired knowledge, and theoretical advances. This alone earns them the rubric, "intelligent" systems.

About one-half of the systems investigated are based in North America. Most of the others are based in the United Kingdom. Great Britain has a long and worthy tradition of information system research and experimentation. Fortunately that interest and experience is being applied to OPACs and bibliographic retrieval systems. In most cases this investigator had access to the system or software for actual use. Demonstrations sufficed in a few cases. Technical documentation and screen illustrations supplied by tne vendor, designer, or researcher responsible for the system/software supported the study. Many published articles describing information system research and retrieval experiments were identified and reviewed. In many cases it was the good fortune of the investigator to visit and converse with key researchers and designers.

It is not the purpose of this report to present a comprehensive overview of advanced OPAC interface and retrieval system developments. Rather, the aim is to be selective enough to illustrate in a systematic way how major limitations surrounding the usability and effective performance of conventional online retrieval systems are being addressed by researchers and designers using scientifically-based design approaches and "intelligent" software. To accomplish this aim some groundwork must be put down and traversed. An illustrated overview of retrieval features and subject searching in today's "second-generation" OPACs is presented in the next section of this report, followed by a section which outlines the problems and shortcomings of conventional OPACs and online information retrieval systems, especially with regard to subject access and retrieval methods. The illustrated review and analysis of intelligent interfaces and retrieval methods is directly preceded by a summary discussion of perspectives, models, and design contributions from the information retrieval research and experiment community. This discussion facilitates a refined expression and grouping of the major issues and design challenges into four major areas: 1) ease of use, orientation, and presentation factors, 2) vocabulary control and correlation factors, 3) more effective system-guided or automatic query formulation and retrieval techniques, and 4) meaningful engagement of the searcher in relevance assessments, query modification/expansion, and the provision of smart navigational, exploration facilities. A final section summarizes recommended design principles and the achievements in intelligent interface and retrieval system design.

## 2. Retrieval Features of Operational, Conventional Info:mation Retrieval Systems, Including Online Catalogs

### Second Generation Online Public Access Catalogs (OPACs)

Most university-based OPACs in North America and many OPACs supplied by the commercial vendors can be characterized as second-generation OPACs (See Figure 1.). These second-generation OPACs with their MARC catalog record databases can be viewed from a functional perspective as special purpose online reference retrieval systems. Some of them even satisfy Cutter's classic objectives for the library catalog:

1. to enable a person to find a book of which either the author, or the title, or the subject is known;

2. to show what the library has by a given author, or on a given subject, or in a given kind of literature;

3. to assist in the choice of a book, as to its edition, or as to its character (literary or topical).[4]

Second-generation OPACs represent a qualitative leap of progress over first-generation online catalogs, the target of much criticism in the professional literature. These are some of the reasons. In first-generation OPACs searching was initiated by derived-key input or by exact word or phrase- matching on at least the first part (left-most) of the word or phrase (as with heading searches in the card catalog). In addition to lacking full bibliographic records and subject access, including any keyword access to titles or subject headings, first-generation online catalogs provided only a single display format, a single unforgiving mode of interaction with the system, and little or nothing in the way of online user assistance. Refining and improving a search in progress, based on an evaluation of intermediate results, was out of the question. Without full records, subject access, authority-based searching with cross references, and meaningful browsing facilities, first-generation online catalogs were understandably criticized as inferior to traditional library catalogs.

**RETRIEVAL FEATURES**

Subject Access via Library of Congress Headings
Keyword Access to a Variety of Fields
Browsing Index or Thesaurus Terms (With Cross References)
Interactive Search Refinement (Boolean Logic, Limiting, Etc.)
Search Term Approximate Matching (Truncation, Wildcards)
Shelf List Scanning

Two or More Dialogue Modes (Menu, Command, Conversational)
Informative Error Messages
Directory-Based Help Facility
Automatic Search/Display Option Suggestions

Full Bibliographic Records Accessible
Multiple Display Format Options (Inc. Non-MARC Labels)
Search Results Output Options (Save, Printing, Sorting)

**Figure 1.** *2nd Generation OPAC Advances*

Today's second-generation online catalogs represent a marriage of the library catalog and conventional online information retrieval (IR) systems familiar to librarians who search online abstracting and indexing databases via DIALOG, BRS, DATASTAR, MEDLINE, etc. Improved card catalog-like "main entry" searching and browsing-by-heading capabilities have been joined with the conventional IR keyword and Boolean searching approaches. Much of the power and flexibility familiar to online database search specialists which enables the "post-coordinating" of search concepts and terminology has been brought to the OPAC searcher. Many online catalogs support the ability to restrict searches to specified record fields, to perform character-masking and/or righthand truncation, and to limit the results by date, language, place of publication, etc. Also, bibliographic records may be viewed and printed in a number of different display formats.

Second-generation online catalogs should be viewed as online bibliographic information retrieval systems. But when compared to conventional keyword/Boolean commercial database search and retrieval systems, these key differences should be kept in mind:

- the online public access catalog must be usable directly by untrained and inexperienced users (online assistance is usually provided to help with the mechanics of searching),

- records in the catalog database lack abstracts, the subject indexing is sparse and the vocabulary is often not representative of current terminology, and

* the catalog database, in covering a library's collection, includes information on a wide variety of disciplines and subject areas.

Designers of second-generation online catalogs have addressed these differences in two ways: 1) by providing card catalog-like precoordinated phrase searching and browsing options (along with keyword/Boolean capabilities), and 2) by providing more and more online user assistance in the form of menus, help displays, suggestive prompts, and informative error messages.

On the other hand, postcoordinated keyword searching on subject-rich fields (e.g., titles, corporate names, series entries, notes, and subject headings) supported in most of today's OPACs can be used to alleviate the twin problems associated with the sparse subject indexing of most library materials and the users' unfamiliarity with the controlled subject indexing vocabularies (e.g., PRECIS, and "LCSH" - Library of Congress Subject Headings).

A library catalog that fulfills Cutter's classic objectives for the catalog in the online environment is a significant accomplishment. It succeeds in at least two ways: users prefer the online catalog to either the card or the COM catalog, and the online catalog is easier to maintain and update than earlier forms. Designing a keyword/Boolean information retrieval system as an online catalog that is easier to learn and easier to use than the conventional, commercial IR systems is also a significant accomplishment. The traditional, well-structured library catalog has been joined with the power and flexibility of conventional IR systems. Appendix 1 provides a detailed list of second- generation OPAC features. A caveat: No operational OPAC has all of the features and capabilities listed. However, most of these features can be found in the better OPACs. The "second-generation OPAC" is a hypothetical construct which nonetheless represents the operational state-of-the-art system.

## Two Approaches to Searching in Online Catalogs

Two fundamentally different search approaches can be found in second-generation OPACs (See Figure 2). It is commonly recognized that keyword, Boolean searching (postcoordinating) is different than precoordinated phrase searching, but the difference I am addressing now cuts deeper across the OPAC interface territory. It is the difference between exact match *Querying* and *Browsing*. These different search approaches are best understood by considering differences that may in fact exist in 1) users' searching objectives, 2) system query specification and query input requirements, and 3) system output displayed to the searcher and subsequent interaction (if any) with the searcher.

I. QUERYING

A. Phrase Matching
(Text strings or controlled vocabulary)
B. Keyword Matching
(Discrete words, with Boolean or proximity formulations)

**Query search requirements: Search aim/criteria known and can be expressed with relative precision and completeness

II. BROWSING

A. Pre-sequenced, linear, inflexible
(Typically, lists of index terms, headings, descriptors, or brief titles)
B. Non-linear, multidirectional, flexible
("navigation", "chain", "bridge", "relational", "serendipitous" browsing)

**Browse search requirements: Search aim/criteria not specific, not known, and/or cannot be expressed in appropriate query/indexing language

**Figure 2.** *Two Online Catalog Search/Access Options*

There are two kinds of query searching: phrase matching and keyword matching. A query consists of a term or terms (e.g., a character, number, word or words, or a phrase) and the specification, sometimes called the query "formulation," which defines how the component terms of the query are to be interpreted or related for matching purposes (e.g., word truncation, Boolean combinations, word adjacency). The matching function of an OPAC is the mechanism through which the retrieval software makes a comparison between index terms which represent documents and query terms to effect retrieval. The matching criteria are specified through the query by the user or applied automatically by the system. Query searching of either kind (often called just "searching," to distinguish it from an OPAC's Browse mode) utilizes an exact matching function on the part of the system, regardless of the manner in which the matching criteria are specified.

In this all or nothing approach, documents (bibliographic records in OPACs) will be retrieved in response to a search only if an exact match of the query is found. The query may consist of a precoordinated phrase (with or without truncation) or a postcoordinated Boolean expression of keywords. In either case the Query search matching requirements are precise and rigid. The process is purely mechanistic. The burden is on the searcher to enter terms that will match the entry (index) terms in the database and to specify appropriate proximity or term relationship logic. Bates criticizes this predominant approach to subject searching for requiring a "perfect pinpoint match on the one best term."[5] No match means no retrieval, as viewers of our silent OPAC screens witness too often. The search may fail (i.e., not identify relevant documents that are in the collection) unless the searcher knows or guesses the exact way the term (word or phrase) appears in the subject index.

In keyword, Boolean queries, the system's matching mechanism makes a binary (yes/no) split of the database into bibliographic records that conform exactly to the requirements of the query, and all the rest. Only the former are retrieved as "hits." Partial or "closest" matching operations are generally not supported in second-generation OPACs and conventional retrieval systems.

Query searching is an appropriate, useful search option when the aim of the search is specific, when the searcher knows precisely what he wants, and when this request can be expressed in the language of the database. Even in subject searching for books or articles on a topic the searcher may know his topic exactly and may be able to express it in the language of the system (e.g., the assigned subject headings or descriptors).

Browsing in online catalogs can take many forms. Typically, the system displays sequenced lists of terms, descriptors, or brief bibliographic records for scanning by the searcher. Lists of index terms are usually presented in alphabetical order. The arrangement of brief citation records may be according to date, and some systems support short record browsing in shelf-list order. Usually the only "navigation" option for browsers is to go backward or forward through the [illegible] in a semi-constrained, linear manner. Cross references, if included, represent a way of jumping out of the sequence and over to related areas of the database. (Hypertext operations, which permit navigation throughout the database in many different directions and the dynamic definition of "related areas and interests", have not been implemented in second- generation OPACs.) Conventional browsing facilities assume the searcher has a vocabulary-selection/negotiation objective to accomplish near the beginning of a search. They assist in identifying the correct form of a term and any related terms. Other forms of browsing, rare in today's OPACs, support related *record/document* discovery through non-linear explorations of the database. They will be illustrated further along in this report.

Browse searching is the most useful and preferred approach when the search aim is not specific (regarding, for example, discipline or topic, type of publication, level of treatment, perspective, etc.), the desired results are not precisely known in advance, or the correct terms for representing the user's query (which may be vague) are not known at the outset. One or more of these circumstances may be present in most subject searching activities. Researchers are coming to believe that some combination of the above circumstances is the most common situation database searchers present to the information retrieval system. More often than not, searchers want to use an OPAC or IR system because they recognize a gap or anomaly in their state of knowledge. Even if they have some sense of what information they need to resolve the anomaly, they can not precisely describe in a search query what they do not yet precisely know or know at all. As Gerrie explains: "A person cannot specify precisely what is needed to solve his anomaly, let alone attempt to couch that need in a way that defines a potential information source. It is reasonable to assume that there may be a mismatch between an information need that cannot be specified and the fundamental requirement of conventional IR systems for an exact logical expression."[6] Pritchard adds that at the Library of Congress the vastness of the collections "leads even knowledgeable researchers to want to just 'see what there is'" in the Library's catalog.[7]

Some designers of online catalogs have recognized these different search requirements and have provided some rudimentary browsing facilities such as scanning index terms in the same alphabetical neighborhood (See Figure 3.). This feature is not intrinsically tied to either phrase matching or keyword matching search facilities. Phrase matching often incorporates index term browsing automatically in the search dialog, whether or not an initial match occurs, but there are exceptions to this. Most keyword match OPACs now offer the searcher an option to browse alphabetical term and descriptor indexes (See Figures 4 and 5).

Most of today's OPACs support both phrase matching and keyword matching query searching. However, these are hybrid systems that do not integrate or link these separate approaches in any useful way during a search. It is usually up to the searcher to choose, through commands or menu selection, one query type or the other. If the system assumes (defaults to) a particular search operation, the user is not informed which specific operation is being carried out on the search terms. Phrase searching, of course, generally assumes a word adjacency, same word order, matching specification. Keyword searching on two or more words requires the specification of a Boolean operator either by the searcher or by the system. Most OPACs supply the Boolean AND as a default operator between words. When a search term like "economic indicators and business cycles" is entered into an OPAC, if not explicitly instructed by the user the *system* automatically will decide precisely how to interpret the statement for matching purposes. The primary decisions to be made include the indexes to be searched and the logic to be applied to specify the word relationships.

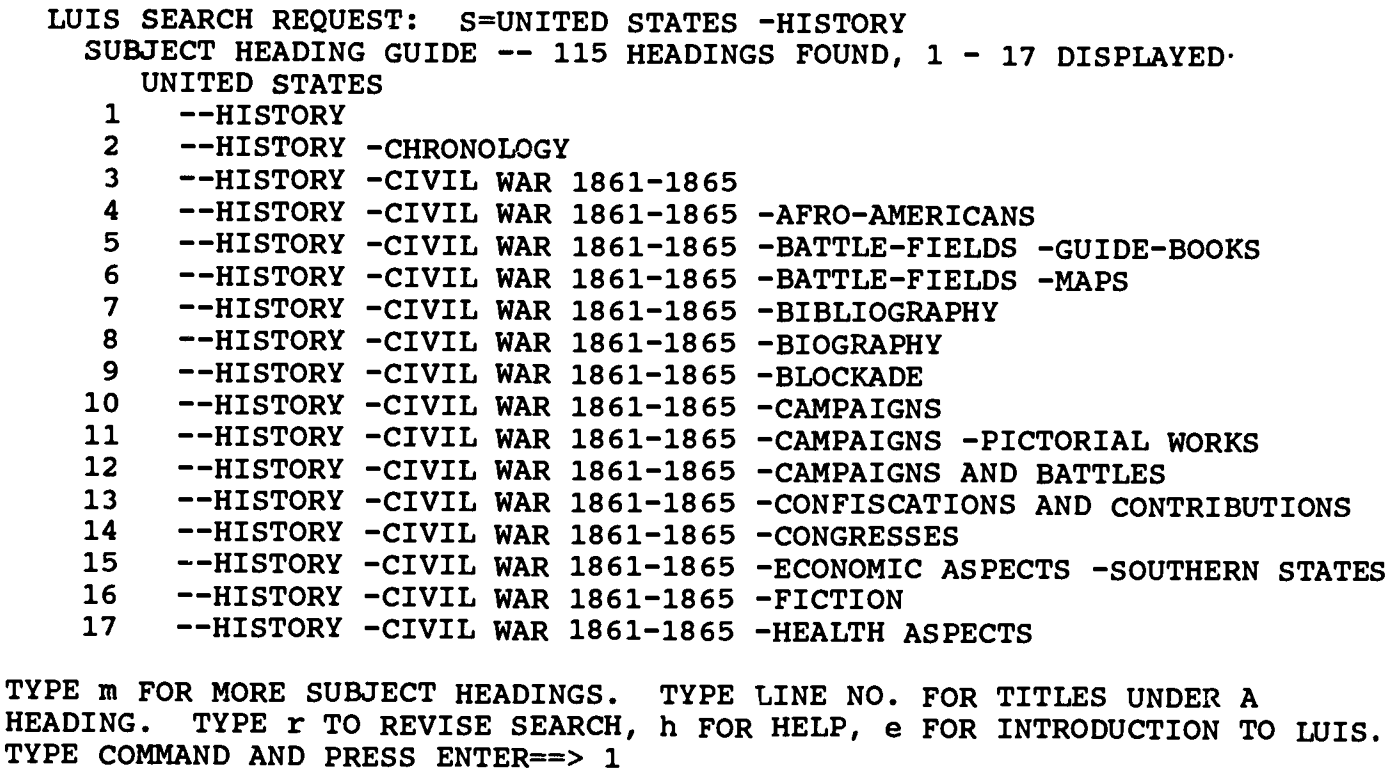
Each type of query search has its advantages, but each may produce very different results even when the same search statement is being processed. This can be a source of confusion for the untrained user who may not understand the relative value of each approach. Such a user may not even know that both ways of searching are supported in a given OPAC, or how to invoke one or the other. Within a given catalog database, one search choice on the same term can produce poor results, another choice, good results. Given these dual, unintegrated search approaches in many OPACs, it is a new responsibility of many librarians to

## RETRIEVAL FEATURES

### *Subject Guide Screen*

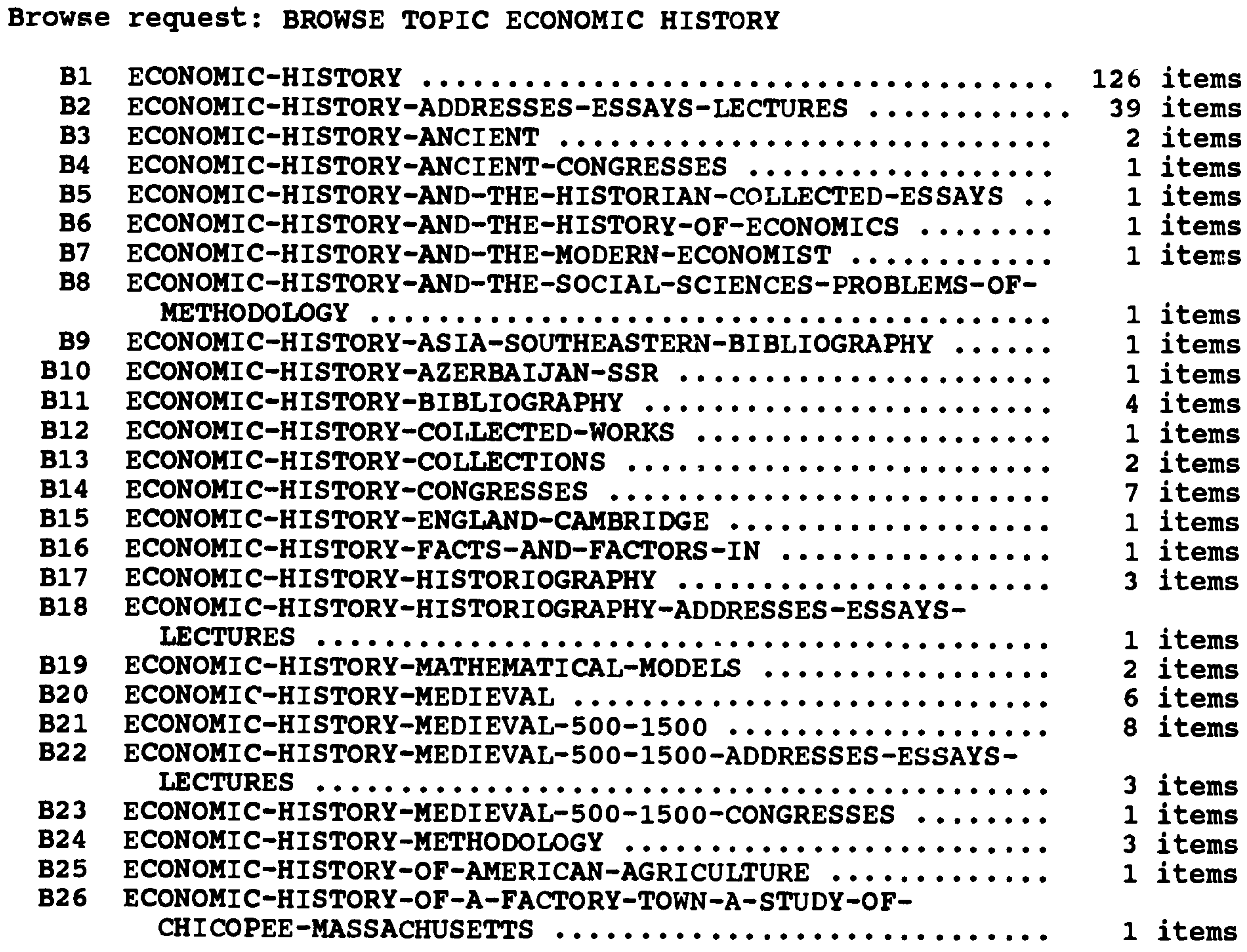
The subject guide screen is displayed when the search statement retrieves more than one subject heading. The guide lists the full text of all the matching subject headings and their subdivisions.

The subject guide screen is illustrated below.



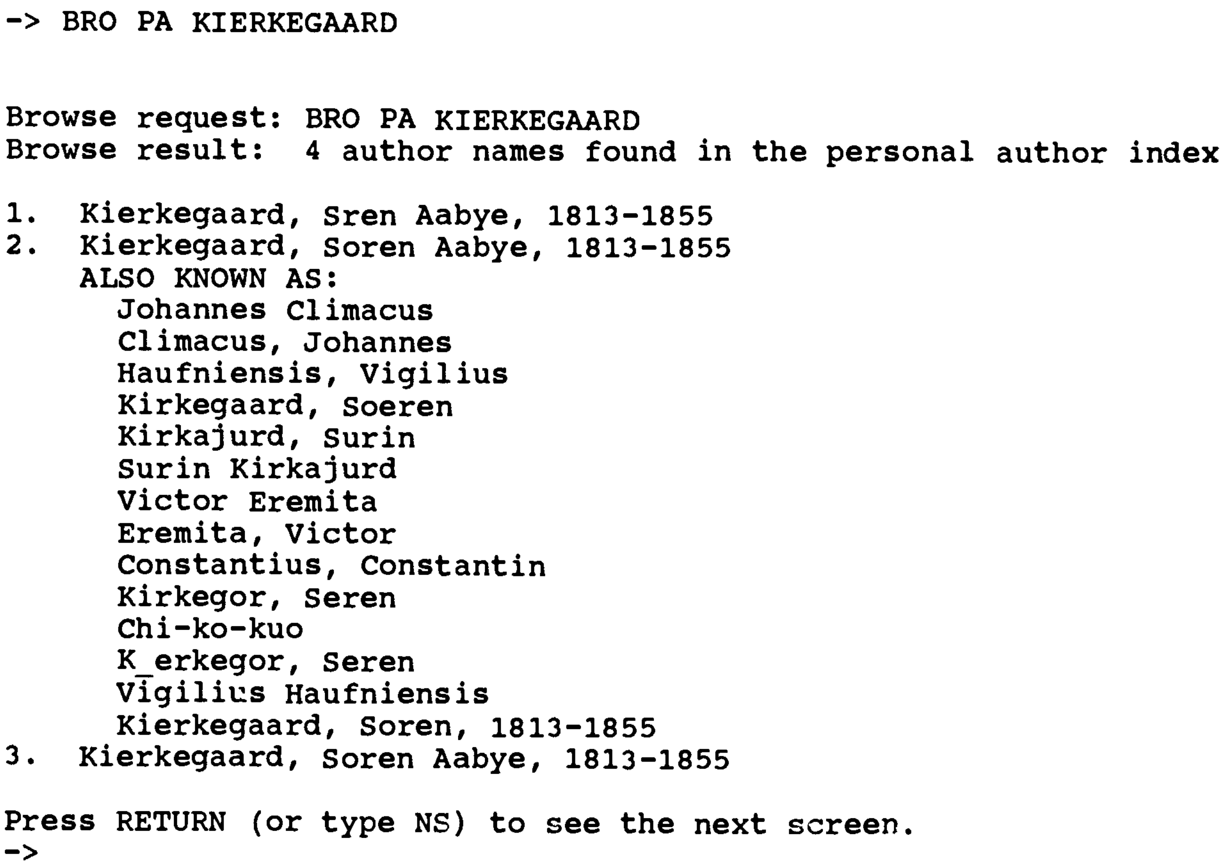
```
   LUIS SEARCH REQUEST:  S=UNITED STATES -HISTORY
     SUBJECT HEADING GUIDE -- 115 HEADINGS FOUND, 1 - 17 DISPLAYED·
        UNITED STATES
      1    --HISTORY
      2    --HISTORY -CHRONOLOGY
      3    --HISTORY -CIVIL WAR 1861-1865
      4    --HISTORY -CIVIL WAR 1861-1865 -AFRO-AMERICANS
      5    --HISTORY -CIVIL WAR 1861-1865 -BATTLE-FIELDS -GUIDE-BOOKS
      6    --HISTORY -CIVIL WAR 1861-1865 -BATTLE-FIELDS -MAPS
      7    --HISTORY -CIVIL WAR 1861-1865 -BIBLIOGRAPHY
      8    --HISTORY -CIVIL WAR 1861-1865 -BIOGRAPHY
      9    --HISTORY -CIVIL WAR 1861-1865 -BLOCKADE
     10    --HISTORY -CIVIL WAR 1861-1865 -CAMPAIGNS
     11    --HISTORY -CIVIL WAR 1861-1865 -CAMPAIGNS -PICTORIAL WORKS
     12    --HISTORY -CIVIL WAR 1861-1865 -CAMPAIGNS AND BATTLES
     13    --HISTORY -CIVIL WAR 1861-1865 -CONFISCATIONS AND CONTRIBUTIONS
     14    --HISTORY -CIVIL WAR 1861-1865 -CONGRESSES
     15    --HISTORY -CIVIL WAR 1861-1865 -ECONOMIC ASPECTS -SOUTHERN STATES
     16    --HISTORY -CIVIL WAR 1861-1865 -FICTION
     17    --HISTORY -CIVIL WAR 1861-1865 -HEALTH ASPECTS

TYPE m FOR MORE SUBJECT HEADINGS.  TYPE LINE NO. FOR TITLES UNDER A
HEADING.  TYPE r TO REVISE SEARCH, h FOR HELP, e FOR INTRODUCTION TO LUIS.
TYPE COMMAND AND PRESS ENTER==> 1
```

**Figure 3.** *NOTIS OPAC Subject Phrase Search Index*



```
Browse request: BROWSE TOPIC ECONOMIC HISTORY

   B1   ECONOMIC-HISTORY ...................................... 126 items
   B2   ECONOMIC-HISTORY-ADDRESSES-ESSAYS-LECTURES ............  39 items
   B3   ECONOMIC-HISTORY-ANCIENT ..............................   2 items
   B4   ECONOMIC-HISTORY-ANCIENT-CONGRESSES ...................   1 items
   B5   ECONOMIC-HISTORY-AND-THE-HISTORIAN-COLLECTED-ESSAYS ..   1 items
   B6   ECONOMIC-HISTORY-AND-THE-HISTORY-OF-ECONOMICS ........   1 items
   B7   ECONOMIC-HISTORY-AND-THE-MODERN-ECONOMIST ............   1 items
   B8   ECONOMIC-HISTORY-AND-THE-SOCIAL-SCIENCES-PROBLEMS-OF-
          METHODOLOGY ........................................   1 items
   B9   ECONOMIC-HISTORY-ASIA-SOUTHEASTERN-BIBLIOGRAPHY ......   1 items
  B10   ECONOMIC-HISTORY-AZERBAIJAN-SSR ......................   1 items
  B11   ECONOMIC-HISTORY-BIBLIOGRAPHY ........................   4 items
  B12   ECONOMIC-HISTORY-COLLECTED-WORKS .....................   1 items
  B13   ECONOMIC-HISTORY-COLLECTIONS .........................   2 items
  B14   ECONOMIC-HISTORY-CONGRESSES ..........................   7 items
  B15   ECONOMIC-HISTORY-ENGLAND-CAMBRIDGE ...................   1 items
  B16   ECONOMIC-HISTORY-FACTS-AND-FACTORS-IN ................   1 items
  B17   ECONOMIC-HISTORY-HISTORIOGRAPHY ......................   3 items
  B18   ECONOMIC-HISTORY-HISTORIOGRAPHY-ADDRESSES-ESSAYS-
          LECTURES ...........................................   1 items
  B19   ECONOMIC-HISTORY-MATHEMATICAL-MODELS .................   2 items
  B20   ECONOMIC-HISTORY-MEDIEVAL ............................   6 items
  B21   ECONOMIC-HISTORY-MEDIEVAL-500-1500 ...................   8 items
  B22   ECONOMIC-HISTORY-MEDIEVAL-500-1500-ADDRESSES-ESSAYS-
          LECTURES ...........................................   3 items
  B23   ECONOMIC-HISTORY-MEDIEVAL-500-1500-CONGRESSES ........   1 items
  B24   ECONOMIC-HISTORY-METHODOLOGY .........................   3 items
  B25   ECONOMIC-HISTORY-OF-AMERICAN-AGRICULTURE .............   1 items
  B26   ECONOMIC-HISTORY-OF-A-FACTORY-TOWN-A-STUDY-OF-
          CHICOPEE-MASSACHUSETTS .............................   1 items
```

**Figure 4.** *Dartmouth College OPAC Subject Browse with Postings Information*



```
-> BRO PA KIERKEGAARD

Browse request: BRO PA KIERKEGAARD
Browse result:  4 author names found in the personal author index

1.  Kierkegaard, Sren Aabye, 1813-1855
2.  Kierkegaard, Soren Aabye, 1813-1855
    ALSO KNOWN AS:
      Johannes Climacus
      Climacus, Johannes
      Haufniensis, Vigilius
      Kirkegaard, Soeren
      Kirkajurd, Surin
      Surin Kirkajurd
      Victor Eremita
      Eremita, Victor
      Constantius, Constantin
      Kirkegor, Seren
      Chi-ko-kuo
      K_erkegor, Seren
      Vigilius Haufniensis
      Kierkegaard, Soren, 1813-1855
3.  Kierkegaard, Soren Aabye, 1813-1855

Press RETURN (or type NS) to see the next screen.
->
```

**Figure 5.** *MELVYL, University of California Name Authority Browse Display*

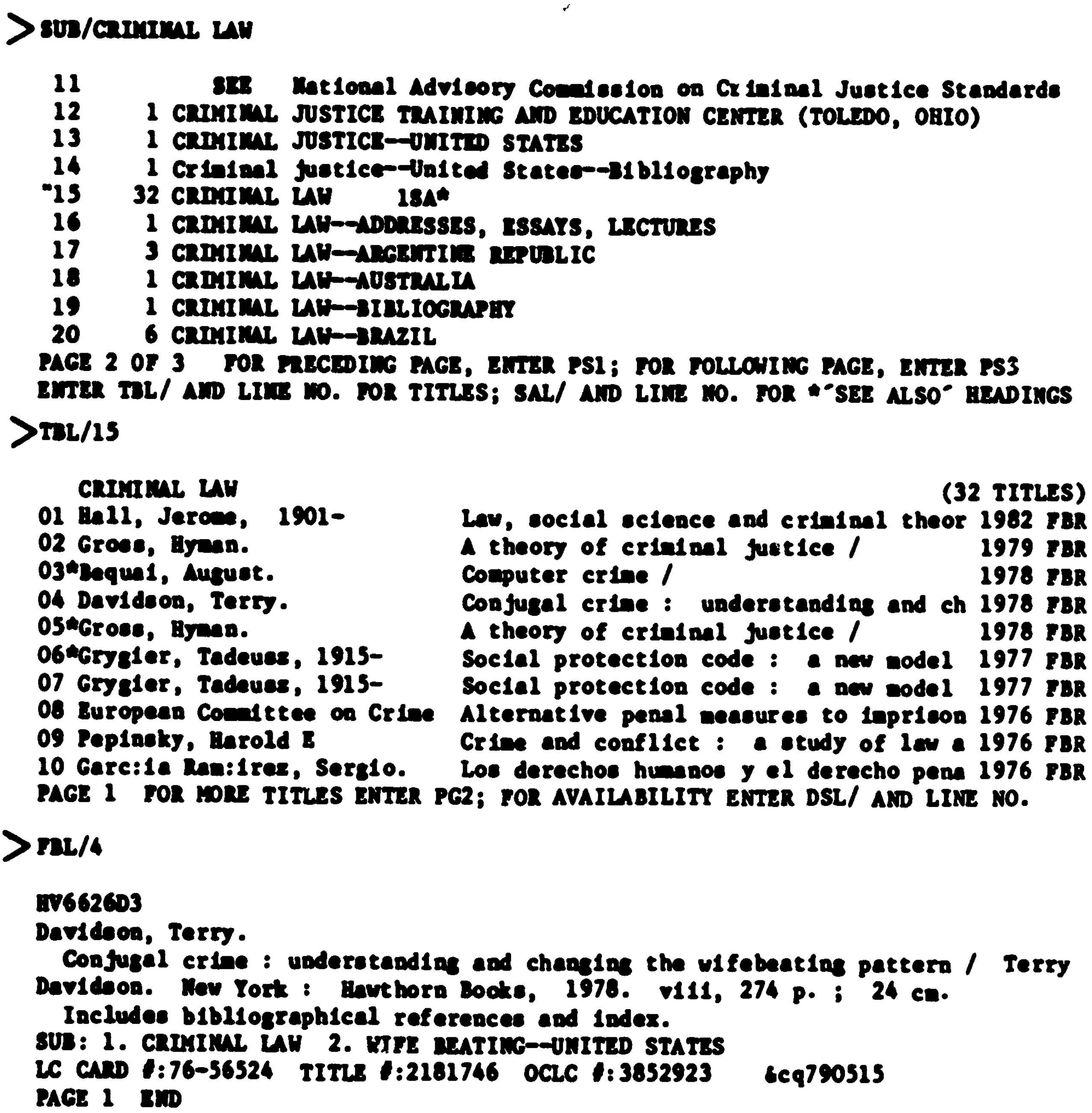
understand their differences and to interpret these differences in a meaningful way for the users of these OPACs.

Figure 6 illustrates a typical search sequence for a phrase-match search using the LCS OPAC at The Ohio State University. The subject search term-/phrase is "criminal law". The search phrase-matched an entry in the subject index, beginning with the left-most word. OPACs providing this search approach usually display a portion of the alphabetical index near the search term, permitting the searcher to browse the index entries. Some OPACs do not display a subject browse list if the search results in a direct match. At least one OPAC displays no subject index in the normal subject phrase-match search process. When provided, this approach to searching is often called "browse" searching to distinguish it from keyword, postcoordinated searching.

The precise phrase depicted in Figure 6 had 32 "hits". After viewing ten short bibliographic records, record number "4" was selected for full display. It is useful to note that this book on criminal law would not have been retrieved had the search been processed as a keyword-in-title search. Searching on the controlled subject vocabulary usually improves recall, the total number of relevant records retrieved from the database for perusal by the searcher. Also, the display of the index entries indicates to the searcher that specific aspects of a broad topic are covered by works in the collection.

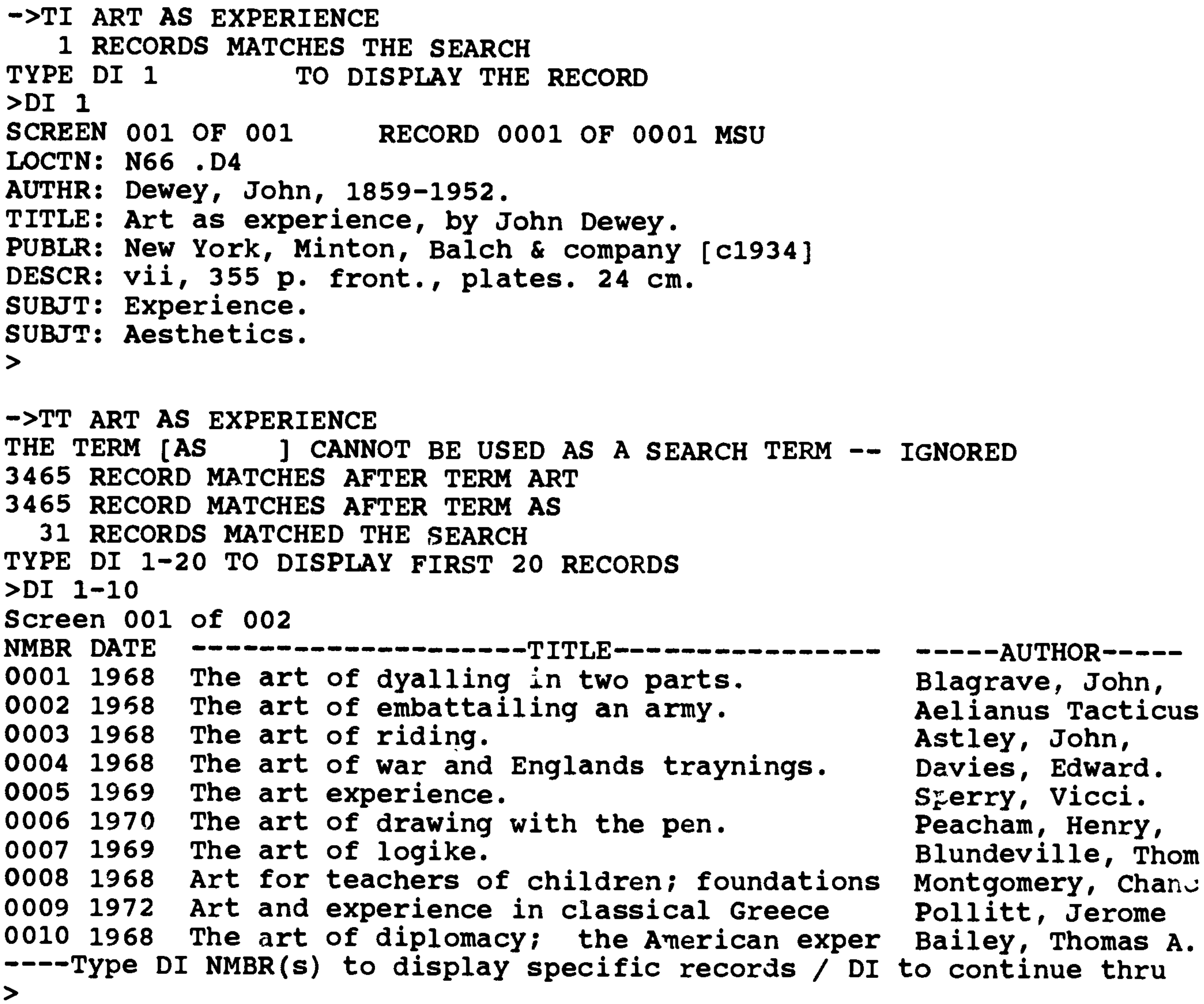
The LCS OPAC at The Ohio State University does not support keyword searching or Boolean combinations of search terms. However, the Minnesota State University System/PALS OPAC supports both types of query searching. Figure 7 illustrates a title phrase search (TI) and a title keyword/Boolean search (TT) on the same expression, "art as experience". Note the very different results! The title phrase search produced a direct hit, a single retrieval. This is very useful for those who know the exact title and want to find the specific item. The second search processed each word in the search statement for a match and then combined them with the Boolean AND operator. Ignoring the stoplisted "as", 31 "matching" records were retrieved, each having the words "art" and "experience" somewhere in their titles, in no particular order or proximity. This is obviously a looser query-document matching logic, useful when the exact title is not known, or when the user is attempting a topical search using keywords to represent relevant concepts.

Figure 8 illustrates more fully the complexity involved when the different approaches to searching are applied to the same search term. Through a seemingly simple change in the search command (TI, TT, TE), very different results are brought to the user, from no matches to 18 matches. Clearly online catalog designers are assuming the user knows how to match the search approach to his search requirements. Research strongly indicates this assumption is unfounded. When the phrase search results in no matches, the PALS system urges the user to "TRY THE ----- TERM SEARCH, IT'S MORE GENERAL". ("Term" search is PALS' keyword search.) In no other way are these different approaches integrated or



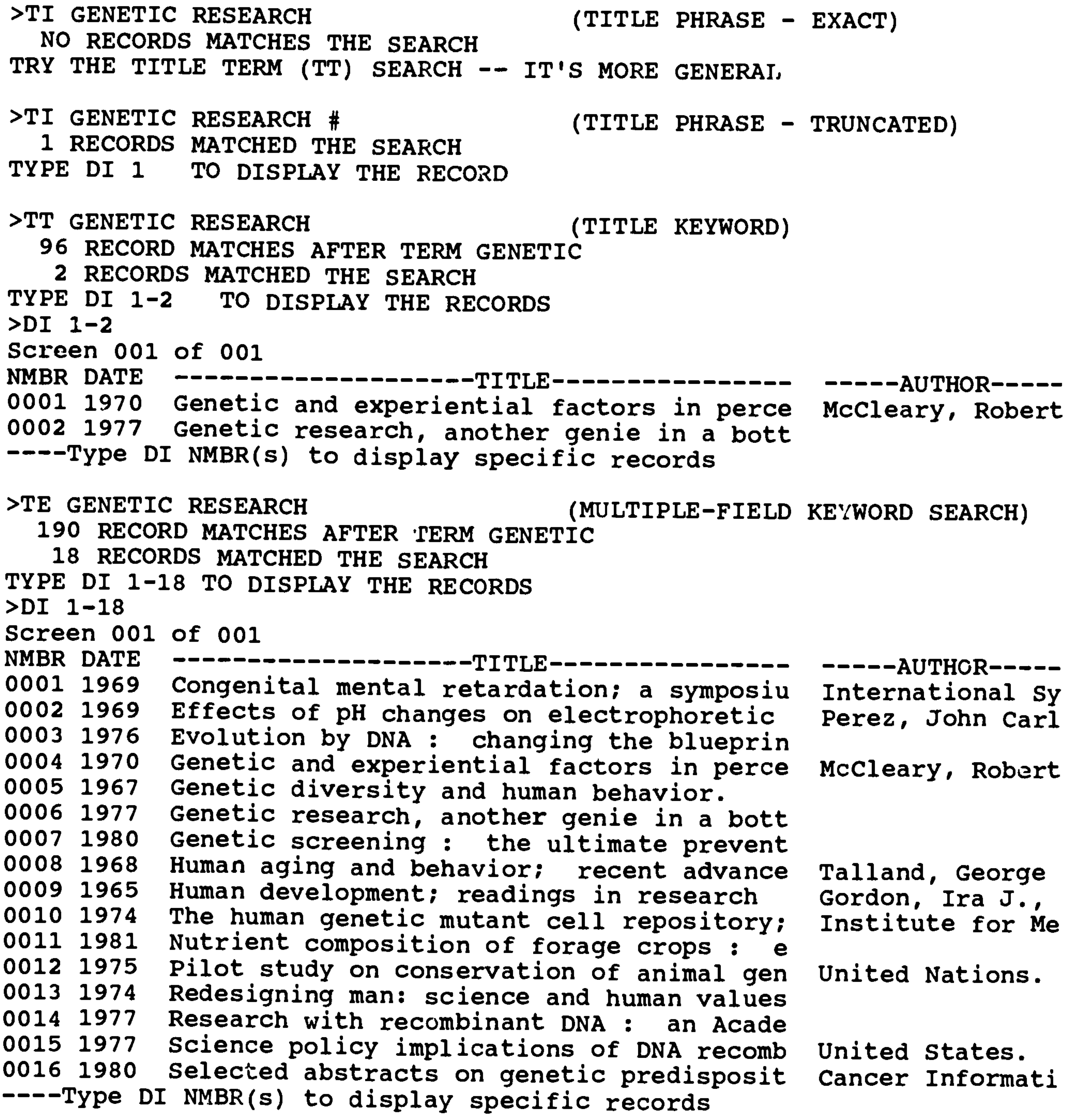
```
>SUB/CRIMINAL LAW

 11          SEE  National Advisory Commission on Criminal Justice Standards
 12       1 CRIMINAL JUSTICE TRAINING AND EDUCATION CENTER (TOLEDO, OHIO)
 13       1 CRIMINAL JUSTICE--UNITED STATES
 14       1 Criminal justice--United States--Bibliography
*15      32 CRIMINAL LAW     1SA*
 16       1 CRIMINAL LAW--ADDRESSES, ESSAYS, LECTURES
 17       3 CRIMINAL LAW--ARGENTINE REPUBLIC
 18       1 CRIMINAL LAW--AUSTRALIA
 19       1 CRIMINAL LAW--BIBLIOGRAPHY
 20       6 CRIMINAL LAW--BRAZIL
PAGE 2 OF 3   FOR PRECEDING PAGE, ENTER PS1; FOR FOLLOWING PAGE, ENTER PS3
ENTER TBL/ AND LINE NO. FOR TITLES; SAL/ AND LINE NO. FOR *'SEE ALSO' HEADINGS
>TBL/15

   CRIMINAL LAW                                                   (32 TITLES)
01 Hall, Jerome,  1901-        Law, social science and criminal theor 1982 FBR
02 Gross, Hyman.               A theory of criminal justice /         1979 FBR
03*Bequai, August.             Computer crime /                       1978 FBR
04 Davidson, Terry.            Conjugal crime :  understanding and ch 1978 FBR
05*Gross, Hyman.               A theory of criminal justice /         1978 FBR
06*Grygier, Tadeusz, 1915-     Social protection code :  a new model  1977 FBR
07 Grygier, Tadeusz, 1915-     Social protection code :  a new model  1977 FBR
08 European Committee on Crime Alternative penal measures to imprison 1976 FBR
09 Pepinsky, Harold E          Crime and conflict :  a study of law a 1976 FBR
10 Garc:ia Ram:irez, Sergio.   Los derechos humanos y el derecho pena 1976 FBR
PAGE 1  FOR MORE TITLES ENTER PG2; FOR AVAILABILITY ENTER DSL/ AND LINE NO.

>FBL/4

HV6626D3
Davidson, Terry.
  Conjugal crime : understanding and changing the wifebeating pattern /  Terry
Davidson.  New York :  Hawthorn Books,  1978.  viii, 274 p. ;  24 cm.
  Includes bibliographical references and index.
SUB: 1. CRIMINAL LAW  2. WIFE BEATING--UNITED STATES
LC CARD #:76-56524  TITLE #:2181746  OCLC #:3852923    &cq790515
PAGE 1  END
```

Figure 6. *Subject Phrase Search on LCS OPAC, The Ohio State University*



```
->TI ART AS EXPERIENCE
   1 RECORDS MATCHES THE SEARCH
TYPE DI 1        TO DISPLAY THE RECORD
>DI 1
SCREEN 001 OF 001     RECORD 0001 OF 0001 MSU
LOCTN: N66 .D4
AUTHR: Dewey, John, 1859-1952.
TITLE: Art as experience, by John Dewey.
PUBLR: New York, Minton, Balch & company [c1934]
DESCR: vii, 355 p. front., plates. 24 cm.
SUBJT: Experience.
SUBJT: Aesthetics.
>

->TT ART AS EXPERIENCE
THE TERM [AS     ] CANNOT BE USED AS A SEARCH TERM -- IGNORED
3465 RECORD MATCHES AFTER TERM ART
3465 RECORD MATCHES AFTER TERM AS
  31 RECORDS MATCHED THE SEARCH
TYPE DI 1-20 TO DISPLAY FIRST 20 RECORDS
>DI 1-10
Screen 001 of 002
NMBR DATE  --------------------TITLE----------------  -----AUTHOR-----
0001 1968  The art of dyalling in two parts.          Blagrave, John,
0002 1968  The art of embattailing an army.           Aelianus Tacticus
0003 1968  The art of riding.                         Astley, John,
0004 1968  The art of war and Englands traynings.     Davies, Edward.
0005 1969  The art experience.                        Sperry, Vicci.
0006 1970  The art of drawing with the pen.           Peacham, Henry,
0007 1969  The art of logike.                         Blundeville, Thom
0008 1968  Art for teachers of children; foundations  Montgomery, Chanc
0009 1972  Art and experience in classical Greece     Pollitt, Jerome
0010 1968  The art of diplomacy;  the American exper  Bailey, Thomas A.
----Type DI NMBR(s) to display specific records / DI to continue thru
>
```

**Figure 7.** *Exact Phrase and Keyword Searches, Minnesota State University System/PALS*



```
>TI GENETIC RESEARCH                       (TITLE PHRASE - EXACT)
  NO RECORDS MATCHES THE SEARCH
TRY THE TITLE TERM (TT) SEARCH -- IT'S MORE GENERAL

>TI GENETIC RESEARCH #                     (TITLE PHRASE - TRUNCATED)
  1 RECORDS MATCHED THE SEARCH
TYPE DI 1   TO DISPLAY THE RECORD

>TT GENETIC RESEARCH                       (TITLE KEYWORD)
  96 RECORD MATCHES AFTER TERM GENETIC
   2 RECORDS MATCHED THE SEARCH
TYPE DI 1-2   TO DISPLAY THE RECORDS
>DI 1-2
Screen 001 of 001
NMBR DATE  -------------------TITLE---------------  -----AUTHOR-----
0001 1970  Genetic and experiential factors in perce  McCleary, Robert
0002 1977  Genetic research, another genie in a bott
----Type DI NMBR(s) to display specific records

>TE GENETIC RESEARCH                       (MULTIPLE-FIELD KEYWORD SEARCH)
  190 RECORD MATCHES AFTER TERM GENETIC
   18 RECORDS MATCHED THE SEARCH
TYPE DI 1-18 TO DISPLAY THE RECORDS
>DI 1-18
Screen 001 of 001
NMBR DATE  -------------------TITLE---------------  -----AUTHOR-----
0001 1969  Congenital mental retardation; a symposiu  International Sy
0002 1969  Effects of pH changes on electrophoretic   Perez, John Carl
0003 1976  Evolution by DNA :  changing the blueprin
0004 1970  Genetic and experiential factors in perce  McCleary, Robert
0005 1967  Genetic diversity and human behavior.
0006 1977  Genetic research, another genie in a bott
0007 1980  Genetic screening :  the ultimate prevent
0008 1968  Human aging and behavior;  recent advance  Talland, George
0009 1965  Human development; readings in research    Gordon, Ira J.,
0010 1974  The human genetic mutant cell repository;  Institute for Me
0011 1981  Nutrient composition of forage crops :  e
0012 1975  Pilot study on conservation of animal gen  United Nations.
0013 1974  Redesigning man: science and human values
0014 1977  Research with recombinant DNA :  an Acade
0015 1977  Science policy implications of DNA recomb  United States.
0016 1980  Selected abstracts on genetic predisposit  Cancer Informati
----Type DI NMBR(s) to display specific records
```

**Figure 8.** *PALS Search Options*

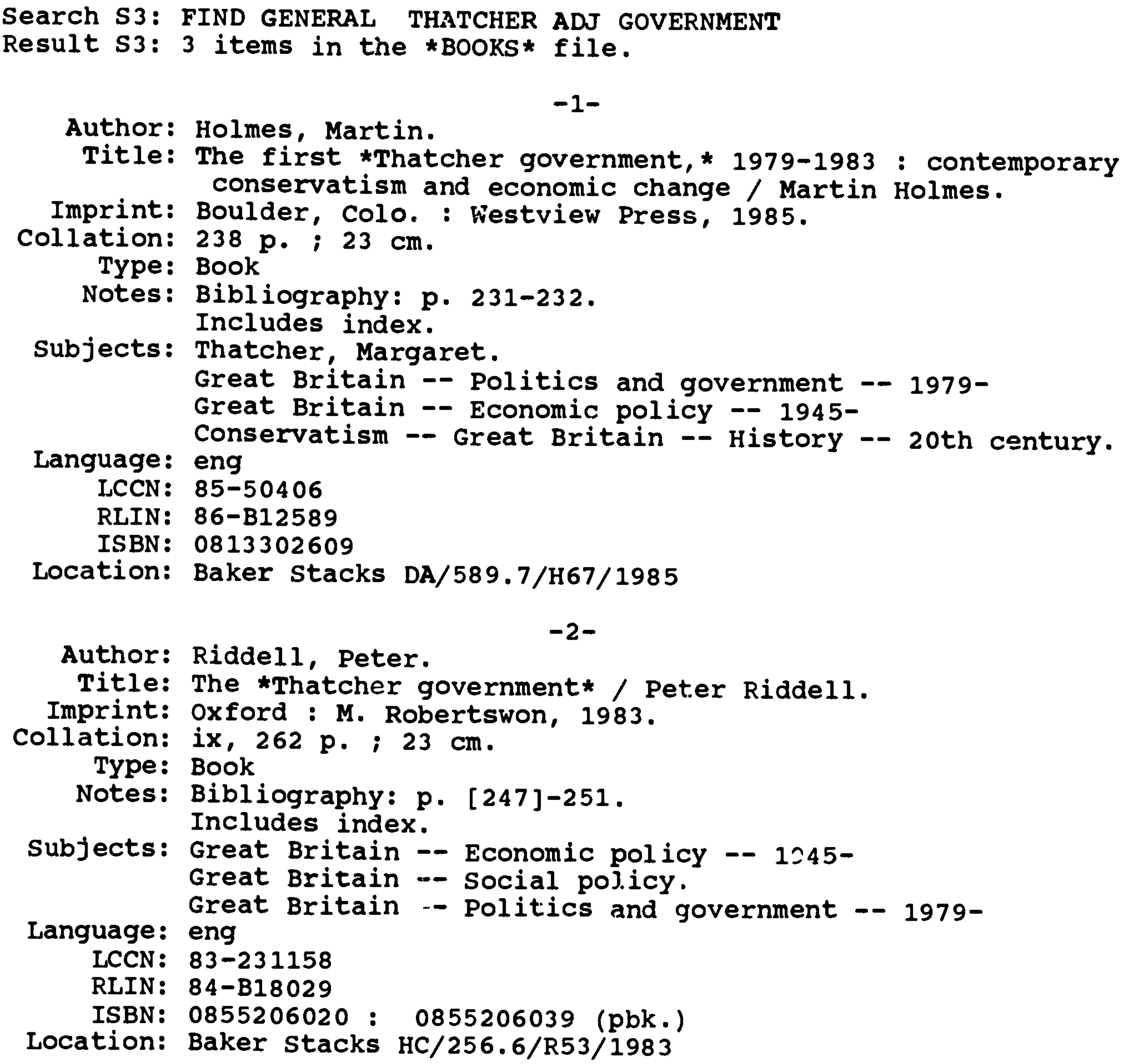
coordinated in the PALS OPAC. Also, the searcher is not guided from successful keyword searches to associated subject headings or call numbers to retrieve additional relevant records. This could be done through an explicit suggestive prompt or semi-automatically triggered by the user's approval.

Figure 9 illustrates the proximity searching feature recently introduced on the keyword-oriented Dartmouth College OPAC. "GENERAL" instructs the system to search several different word indexes (e.g., subject, title, notes), and "ADJ" specifies an adjacency, same order match on the two search words, "THATCHER' and "GOVERNMENT." In the absence of the ADJ specification, the system would process this search as a Boolean AND query. In either case, only an exact match of the *query* will result in a successful retrieval. "FIND GENERAL THATCHER GOVERNMENT" would retrieve these two citations and any others having both "THATCHER" and "GOVERNMENT" anywhere in the record because that is exactly what the query specifies. On the other hand, if a record in the database contained the word "Thatcher" (the OPAC is upper/lower case- insensitive) but not "Government," and contained the word "Politics" it would not be retrieved by this Boolean AND query even though it would probably be relevant. Transaction logs indicate most users' subject searches entered in this catalog contain no more than two words.

## Easier Subject Searching

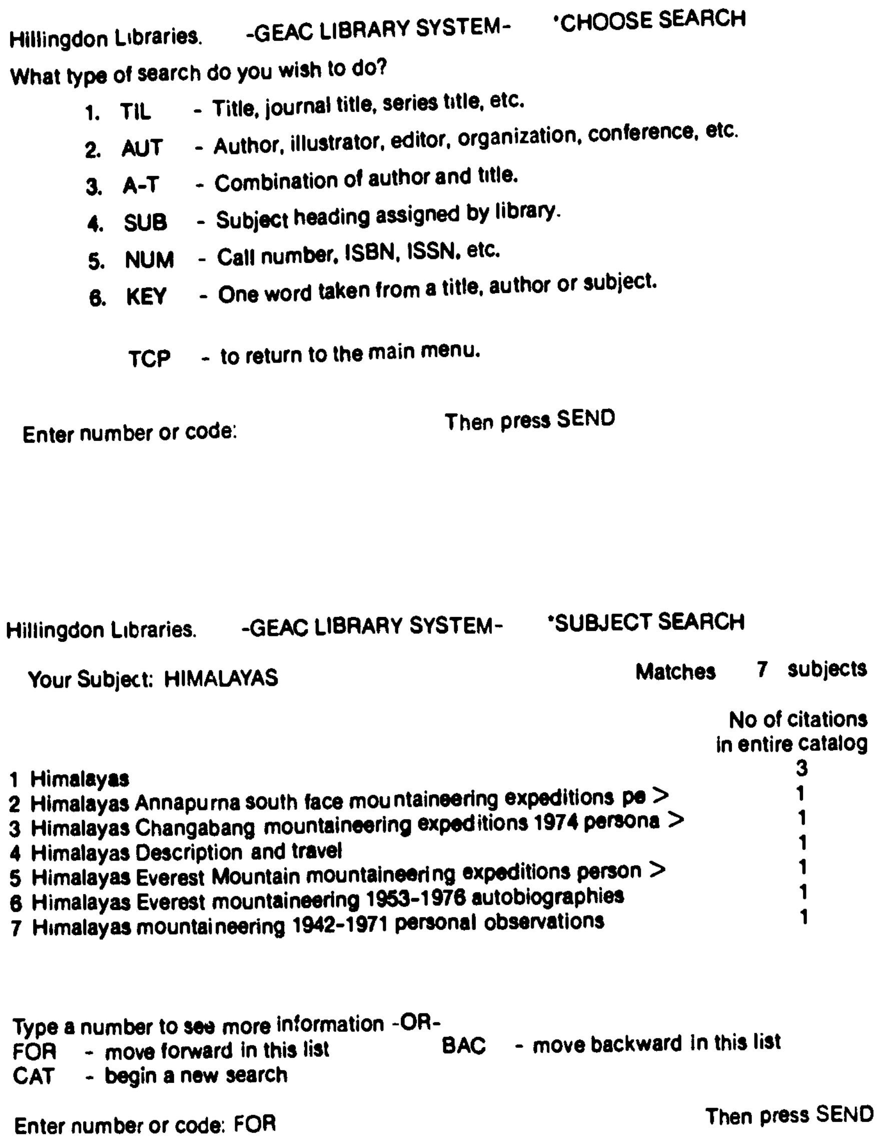
Recent improvements to user interface software have made subject searching in OPACs easier, if not more effective. However, these improvements are independent of, and have had no impact on, the query matching requirements and retrieval methods previously described. Yet, OPAC design is clearly moving in the right direction: namely, to a greater recognition of the needs and difficulties of untrained or casual users of online catalogs. Figures 10-13 illustrate how new interface design features are making it easier to select and enter a subject search in OPACs. Difficult to learn command languages with their complex syntax requirements have largely disappeared from the OPAC interface. They have been replaced by straightforward menu selections (Figures 10 & 11), prompted search term entry (Figure 12), and newer approaches such as graphic entry devices (window search logic boxes and onscreen search worksheets), and query-by-example or query-by-form (Figure 13). With query-by- example a retrieved record can serve as a model template to initiate, through editing, a new or modified subject search. A query-by-form template is a prompted, preformatted search entry, citation-like, form just needing data in chosen fields to initiate a search. The fields and field labels can be customized by the library staff. Its value lies in the intuitive association it triggers between query expressions and document representations (citations) in the catalog database.

In her landmark study of subject access problems and opportunities in OPACs, Markey pleads for a "forgiving and simple implementation of Boolean search capabilities."[8] This is now being achieved through simplified query



```
Search S3: FIND GENERAL  THATCHER ADJ GOVERNMENT
Result S3: 3 items in the *BOOKS* file.

                                   -1-
    Author: Holmes, Martin.
     Title: The first *Thatcher government,* 1979-1983 : contemporary
             conservatism and economic change / Martin Holmes.
   Imprint: Boulder, Colo. : Westview Press, 1985.
 Collation: 238 p. ; 23 cm.
      Type: Book
     Notes: Bibliography: p. 231-232.
            Includes index.
  Subjects: Thatcher, Margaret.
            Great Britain -- Politics and government -- 1979-
            Great Britain -- Economic policy -- 1945-
            Conservatism -- Great Britain -- History -- 20th century.
  Language: eng
      LCCN: 85-50406
      RLIN: 86-B12589
      ISBN: 0813302609
  Location: Baker Stacks DA/589.7/H67/1985

                                   -2-
    Author: Riddell, Peter.
     Title: The *Thatcher government* / Peter Riddell.
   Imprint: Oxford : M. Robertswon, 1983.
 Collation: ix, 262 p. ; 23 cm.
      Type: Book
     Notes: Bibliography: p. [247]-251.
            Includes index.
  Subjects: Great Britain -- Economic policy -- 1945-
            Great Britain -- Social policy.
            Great Britain -- Politics and government -- 1979-
  Language: eng
      LCCN: 83-231158
      RLIN: 84-B18029
      ISBN: 0855206020 :  0855206039 (pbk.)
  Location: Baker Stacks HC/256.6/R53/1983
```

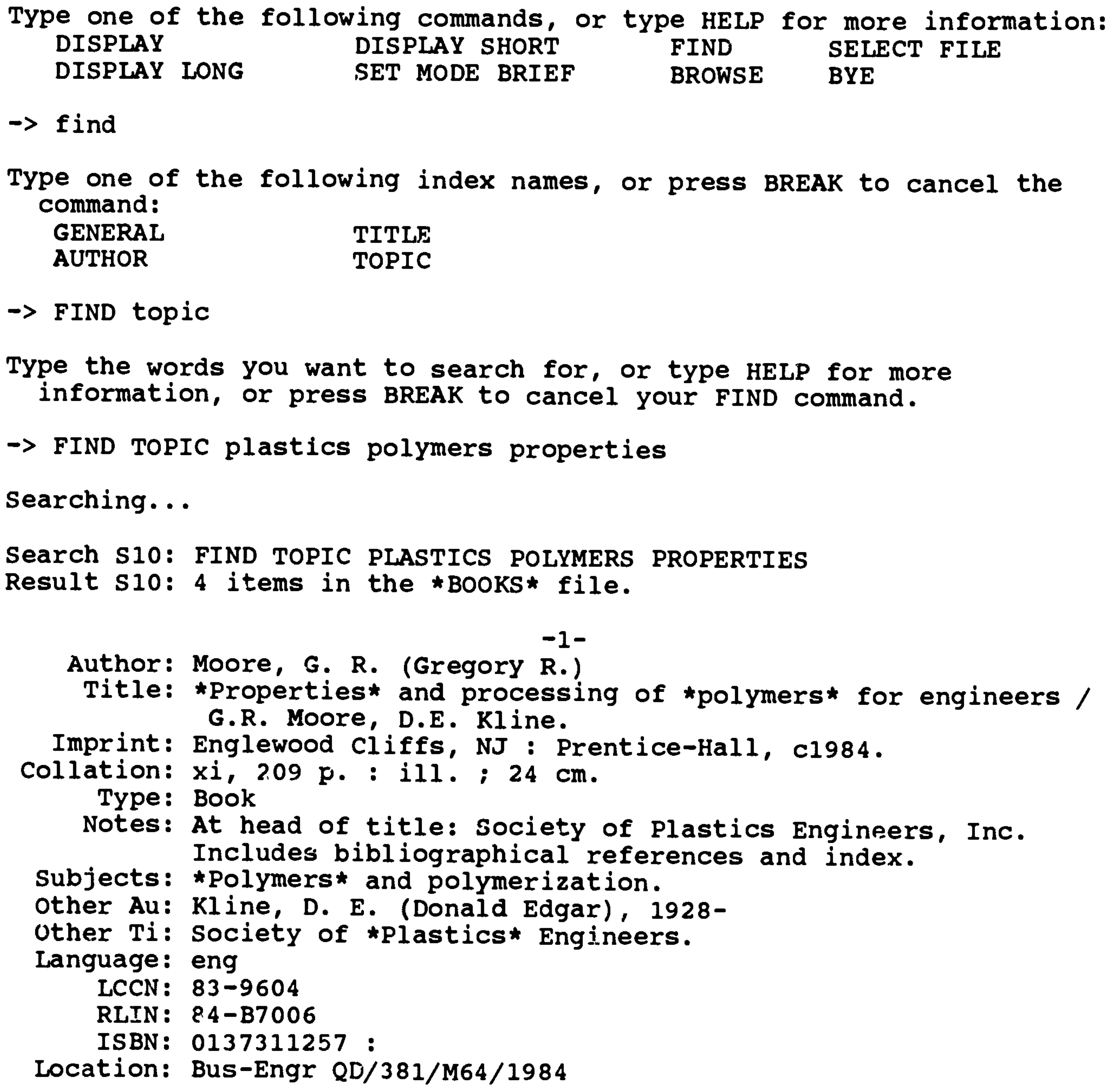
**Figure 9.** *Dartmouth College OPAC Adjacency Search*



```
Hillingdon Libraries.     -GEAC LIBRARY SYSTEM-     *CHOOSE SEARCH
What type of search do you wish to do?
        1.  TIL    - Title, journal title, series title, etc.
        2.  AUT    - Author, illustrator, editor, organization, conference, etc.
        3.  A-T    - Combination of author and title.
        4.  SUB    - Subject heading assigned by library.
        5.  NUM    - Call number, ISBN, ISSN, etc.
        6.  KEY    - One word taken from a title, author or subject.

            TCP    - to return to the main menu.

 Enter number or code:                    Then press SEND
```

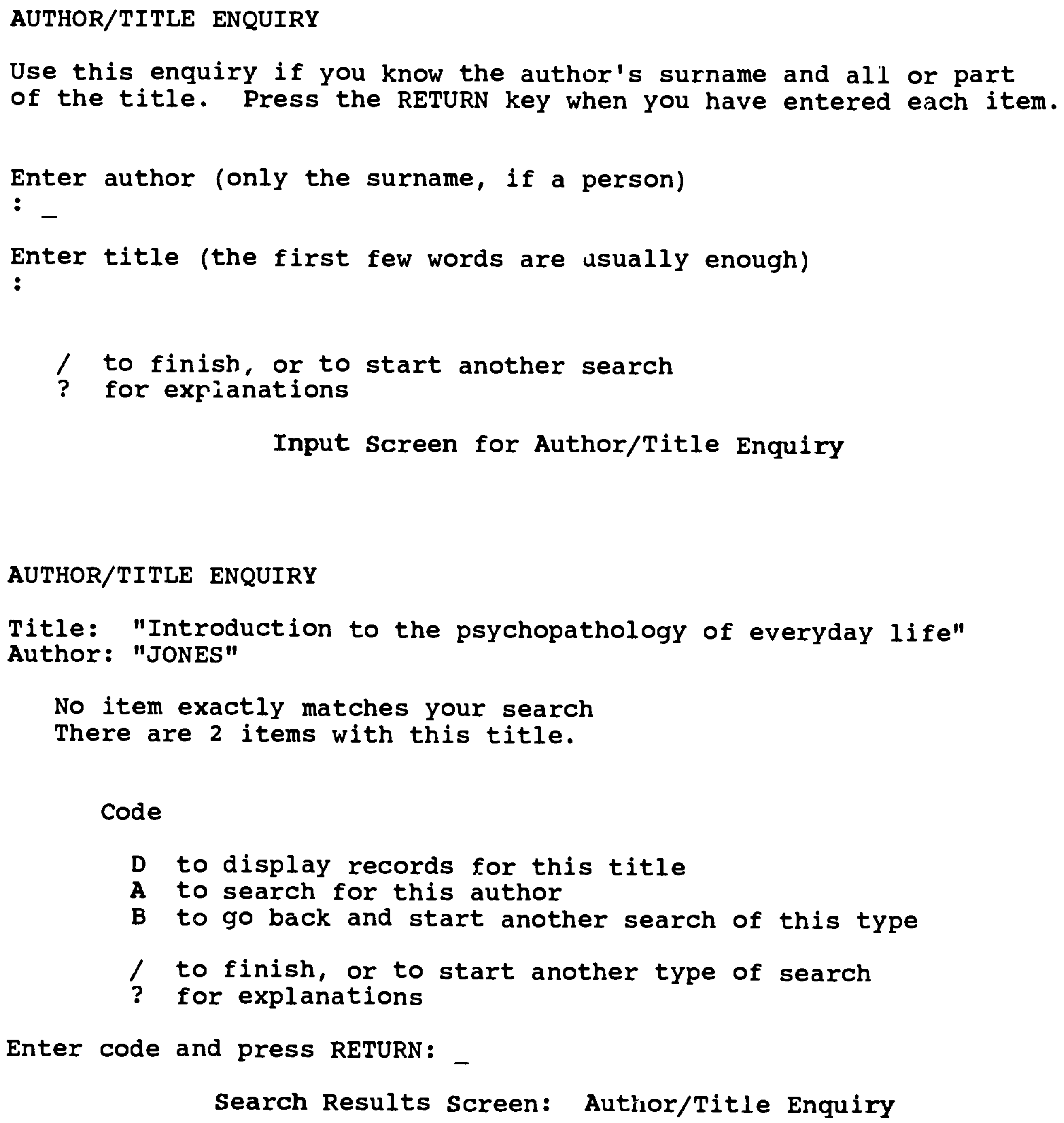
```
Hillingdon Libraries.     -GEAC LIBRARY SYSTEM-     *SUBJECT SEARCH
  Your Subject: HIMALAYAS                                Matches    7  subjects

                                                              No of citations
                                                             in entire catalog
1 Himalayas                                                          3
2 Himalayas Annapurna south face mountaineering expeditions pe >     1
3 Himalayas Changabang mountaineering expeditions 1974 persona >     1
4 Himalayas Description and travel                                   1
5 Himalayas Everest Mountain mountaineering expeditions person >     1
6 Himalayas Everest mountaineering 1953-1976 autobiographies         1
7 Himalayas mountaineering 1942-1971 personal observations           1


Type a number to see more information -OR-
FOR   - move forward in this list          BAC   - move backward in this list
CAT   - begin a new search

Enter number or code: FOR                                   Then press SEND
```

**Figure 10.** *GEAC Search Menu and Subject Browse Screens*



```
Type one of the following commands, or type HELP for more information:
   DISPLAY             DISPLAY SHORT       FIND      SELECT FILE
   DISPLAY LONG        SET MODE BRIEF      BROWSE    BYE

-> find

Type one of the following index names, or press BREAK to cancel the
  command:
   GENERAL             TITLE
   AUTHOR              TOPIC

-> FIND topic

Type the words you want to search for, or type HELP for more
  information, or press BREAK to cancel your FIND command.

-> FIND TOPIC plastics polymers properties

Searching...

Search S10: FIND TOPIC PLASTICS POLYMERS PROPERTIES
Result S10: 4 items in the *BOOKS* file.

                          -1-
    Author: Moore, G. R. (Gregory R.)
     Title: *Properties* and processing of *polymers* for engineers /
             G.R. Moore, D.E. Kline.
   Imprint: Englewood Cliffs, NJ : Prentice-Hall, c1984.
 Collation: xi, 209 p. : ill. ; 24 cm.
      Type: Book
     Notes: At head of title: Society of Plastics Engineers, Inc.
            Includes bibliographical references and index.
  Subjects: *Polymers* and polymerization.
  Other Au: Kline, D. E. (Donald Edgar), 1928-
  Other Ti: Society of *Plastics* Engineers.
  Language: eng
      LCCN: 83-9604
      RLIN: 84-B7006
      ISBN: 0137311257 :
  Location: Bus-Engr QD/381/M64/1984
```

**Figure 11.** *Dartmouth College OPAC Search in Progress*



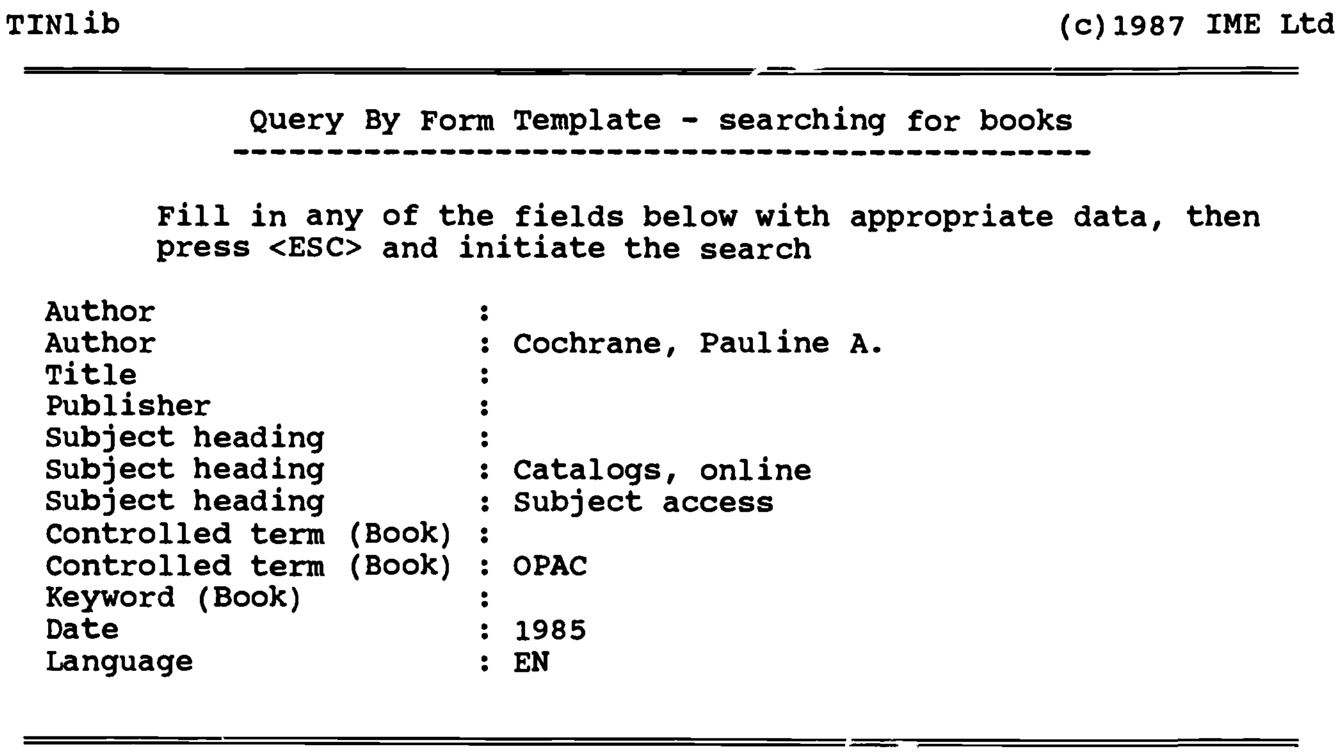
```
AUTHOR/TITLE ENQUIRY

Use this enquiry if you know the author's surname and all or part
of the title.  Press the RETURN key when you have entered each item.


Enter author (only the surname, if a person)
: _

Enter title (the first few words are usually enough)
:


   /  to finish, or to start another search
   ?  for explanations
```

Input Screen for Author/Title Enquiry

```
AUTHOR/TITLE ENQUIRY

Title:  "Introduction to the psychopathology of everyday life"
Author: "JONES"

   No item exactly matches your search
   There are 2 items with this title.


      Code

        D  to display records for this title
        A  to search for this author
        B  to go back and start another search of this type

        /  to finish, or to start another type of search
        ?  for explanations

Enter code and press RETURN: _
```

Search Results Screen: Author/Title Enquiry

**Figure 12.** *SWALCAP's Libertas OPAC*



```
TINlib                                                    (c)1987 IME Ltd

             Query By Form Template - searching for books
             ----------------------------------------------

       Fill in any of the fields below with appropriate data, then
       press <ESC> and initiate the search

   Author                   :
   Author                   : Cochrane, Pauline A.
   Title                    :
   Publisher                :
   Subject heading          :
   Subject heading          : Catalogs, online
   Subject heading          : Subject access
   Controlled term (Book)   :
   Controlled term (Book)   : OPAC
   Keyword (Book)           :
   Date                     : 1985
   Language                 : EN
```

**Figure 13.** *TINlib's Query-By-Example, Query-By-Form Template*

interfaces which do not require explicit entry of commands or Boolean operators. When a phrase search results in 0 hits, the PALS OPAC (Figure 8) prompts the searcher to enter a "TE" keyword, Boolean search to achieve broader results. Only the search qualifier needs to be changed (from "SU" to "TE"). The Dartmouth College OPAC requires the user to type the desired command word, but it is spelled out on the screen in the list of options. The Boolean AND is assumed between search words when no operator is entered by the searcher.

In the TINlib query-by-form option the data fields are ANDed together in the implicit specification of the query. However, the choice of Boolean logic to be applied between the same or different fields can be customized. Most of these OPACs permit the logical operator assigned by the system to be replaced with another operator explicitly entered by the searcher.

Subject term or heading browsing in OPACs has improved somewhat through the provision of clearer, better labelled displays and the addition of postings data (number of citations which match each term). Some OPACs are incorporating "see" and "see also" cross references in these term browsing lists, and shelf list browsing is becoming a common feature although it is not usually prompted by or linked to term or citation browsing activities.

The implicit or automatic specification and postcoordination of search statements into formal queries is becoming both more flexible and "smarter." Markey's research indicated that "postcoordination features of online catalogs need to automatically or explicitly assist searchers in the selection of searching vocabulary and the combination of terms."[9] More and more the scope of subject keyword or "topic" searches is being automatically extended to include matches in the title index as well as the subject headings index. Some OPACs also search the notes and corporate name field indexes when executing a subject search. This multifield subject searching does not have to be specified in advance by the searcher through a complex query syntax. Analysis of OPAC transaction logs has shown that this widening of the subject search target and the implicit choice of the AND logic rather than adjacency to connect search words improves recall (number of relevant documents retrieved), with little or no decrease in search precision (absence of unwanted documents in the retrieval results), when *MARC* bibliographic records are the target of OPAC subject searching. Perhaps most satisfying to searchers, far fewer "0-hit" ("No matches found") search results occur. As a result of an unplanned accident when the database was loaded, the GEAC OPAC shown in Figure 10 includes PRECIS subject headings from the 690 UK MARC field along with the planned entries from the Library of Congress Subject Headings (LCSH) 650 and 651 fields. Users immediately expressed favor with this unplanned development which improved chances for achieving matches in their subject searching. This simple accident enabled matches to occur on many search words differing only in British or American spelling. Both subject vocabularies will be retained, indexed, and mixed in browsing displays in this OPAC. The OPAC subject searching process in all its varieties has been improved through better displays of retrieved citations. From the use of English language field labels to the highlighting of matched terms, citation displays have been made more comprehensible and more informative. Matched query terms appearing in retrieved bibliographic records can be highlighted through reverse video or lightbar techniques, or through the use of special character or graphic markers (See Figure 11). This feature shows the conceptual context of a search term, and promotes an understanding of catalog record structure and system matching principles. On a negative note, the improved labelling of subject headings and call numbers in displayed citations does not yet include indications of their special *collection* role or suggestions that additional potentially relevant documents are linked to (collocated by) these subject descriptors.

Online user assistance displays and messages designed to help users conduct subject searches and to postcoordinate search terms have improved dramatically in OPACs. In this area, OPAC design leads commercial retrieval system design. Searchers may be asked to add a word (usually for a Boolean AND operation) or to enter limiting information such as a range of dates.

OPAC error messages are usually specific, helpful, and informative:

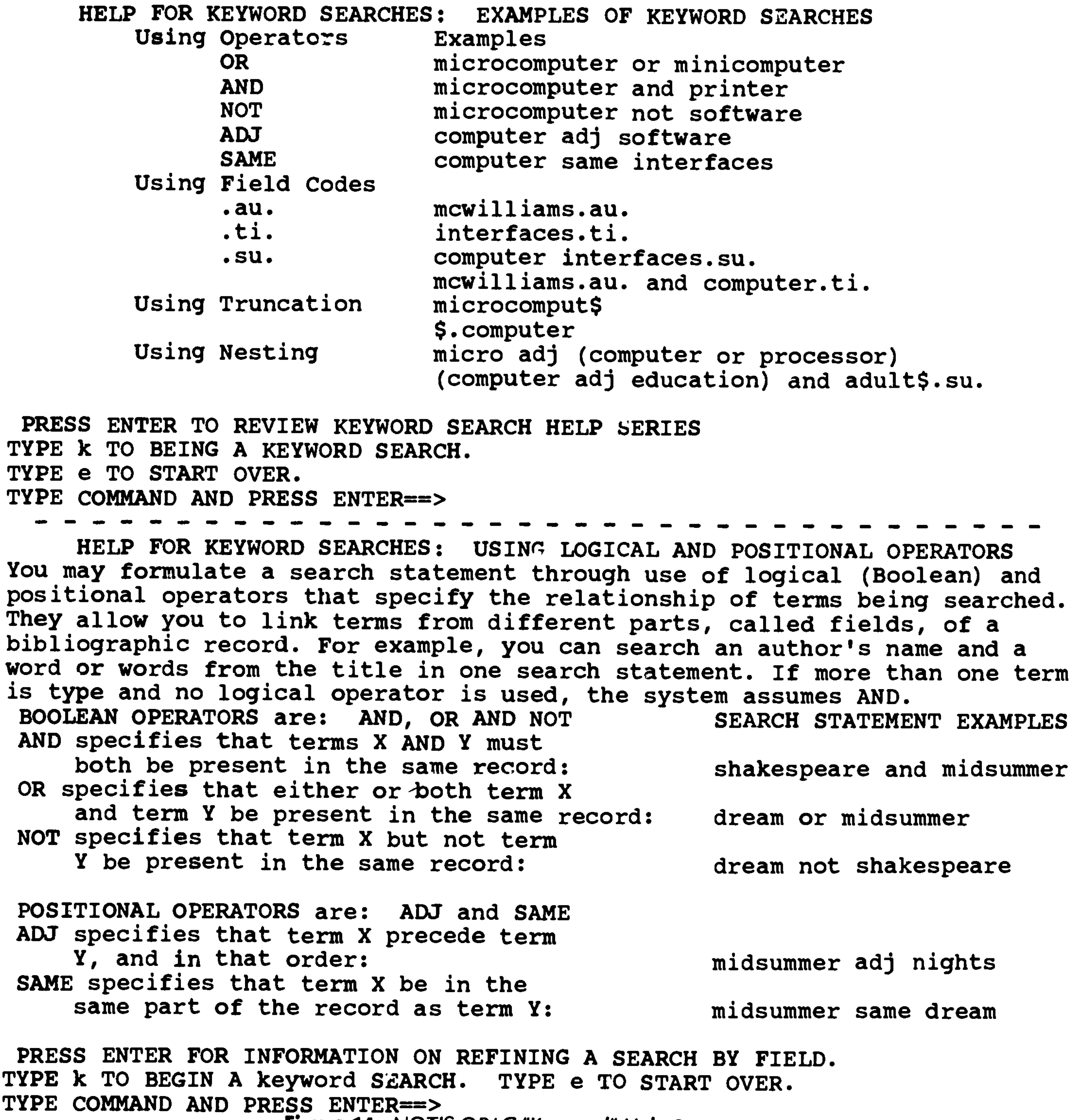
> *"Your BROWSE command does not contain a browsable index name. Please type HELP for more information." (MELVYL)*

Help is frequently context-sensitive, requiring no more than the simple request by the user to bring specific, point-of-need assistance. Specific, addressable help displays are available to assist with the mechanics of searching (See Figure 14) and operation of the local OPAC. Explicit messages and menus embedded in the routine search interaction displays explain things like what the system is doing, what to do next, what options are available at any stage of the search process, and, in some cases, what may be done to improve the search process to produce better results (See Figure 15).

## Conventional, Second-Generation OPAC Subject Searching Summarized

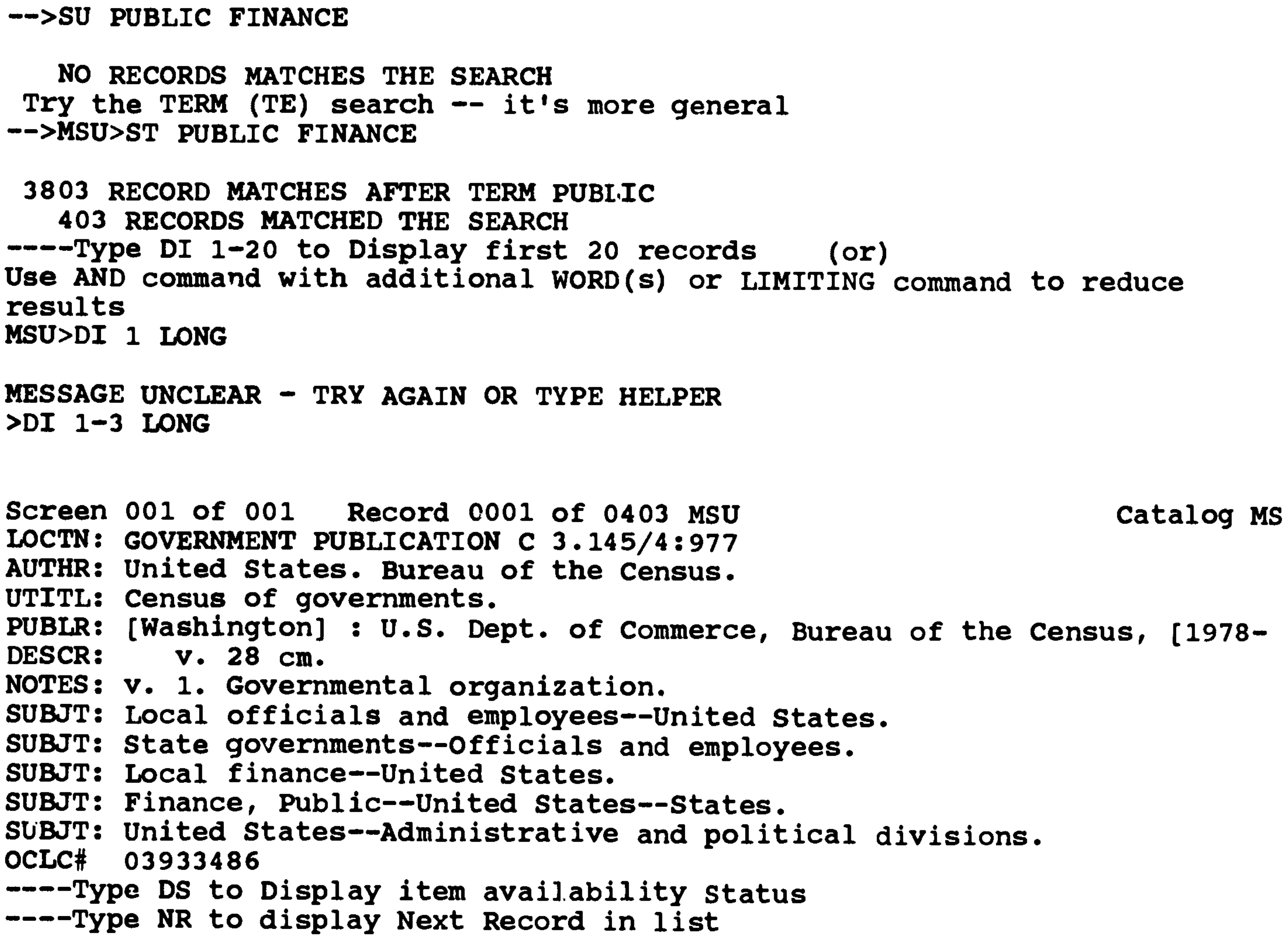
Subject searching in today's largely MARC-based OPACs is supported by a variety of exact match query methods and browsing facilities. Exact match query searches are basically of two types: 1) phrase searching on precoordinated subject headings, and 2) keyword searches on cataloging or indexing data in the bibliographic record (e.g., titles, notes, subject headings and their subdivisions). Keyword subject searches may be formulated as Boolean expressions to indicate the desired relationship between search terms, and some OPACs also permit explicit proximity designations in the formulation of the query. The trend is clear: most OPACs support both types of exact match searching.

The syntax and mechanics of entering subject searches have been simplified, and the task of precise search statement/query formulation has been delegated to the system software. Choice of subject query type is selected from



```
        HELP FOR KEYWORD SEARCHES:  EXAMPLES OF KEYWORD SEARCHES
            Using Operators       Examples
                  OR              microcomputer or minicomputer
                  AND             microcomputer and printer
                  NOT             microcomputer not software
                  ADJ             computer adj software
                  SAME            computer same interfaces
            Using Field Codes
                  .au.            mcwilliams.au.
                  .ti.            interfaces.ti.
                  .su.            computer interfaces.su.
                                  mcwilliams.au. and computer.ti.
            Using Truncation      microcomput$
                                  $.computer
            Using Nesting         micro adj (computer or processor)
                                  (computer adj education) and adult$.su.

    PRESS ENTER TO REVIEW KEYWORD SEARCH HELP SERIES
   TYPE k TO BEING A KEYWORD SEARCH.
   TYPE e TO START OVER.
   TYPE COMMAND AND PRESS ENTER==>
    - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
        HELP FOR KEYWORD SEARCHES:  USING LOGICAL AND POSITIONAL OPERATORS
   You may formulate a search statement through use of logical (Boolean) and
   positional operators that specify the relationship of terms being searched.
   They allow you to link terms from different parts, called fields, of a
   bibliographic record. For example, you can search an author's name and a
   word or words from the title in one search statement. If more than one term
   is type and no logical operator is used, the system assumes AND.
    BOOLEAN OPERATORS are:  AND, OR AND NOT          SEARCH STATEMENT EXAMPLES
    AND specifies that terms X AND Y must
        both be present in the same record:          shakespeare and midsummer
    OR specifies that either or both term X
        and term Y be present in the same record:    dream or midsummer
    NOT specifies that term X but not term
        Y be present in the same record:             dream not shakespeare

    POSITIONAL OPERATORS are:  ADJ and SAME
    ADJ specifies that term X precede term
        Y, and in that order:                        midsummer adj nights
    SAME specifies that term X be in the
        same part of the record as term Y:           midsummer same dream

    PRESS ENTER FOR INFORMATION ON REFINING A SEARCH BY FIELD.
   TYPE k TO BEGIN A keyword SEARCH.  TYPE e TO START OVER.
   TYPE COMMAND AND PRESS ENTER==>
```

**Figure 14.** *NOTIS OPAC "Keyword" Help Screens*



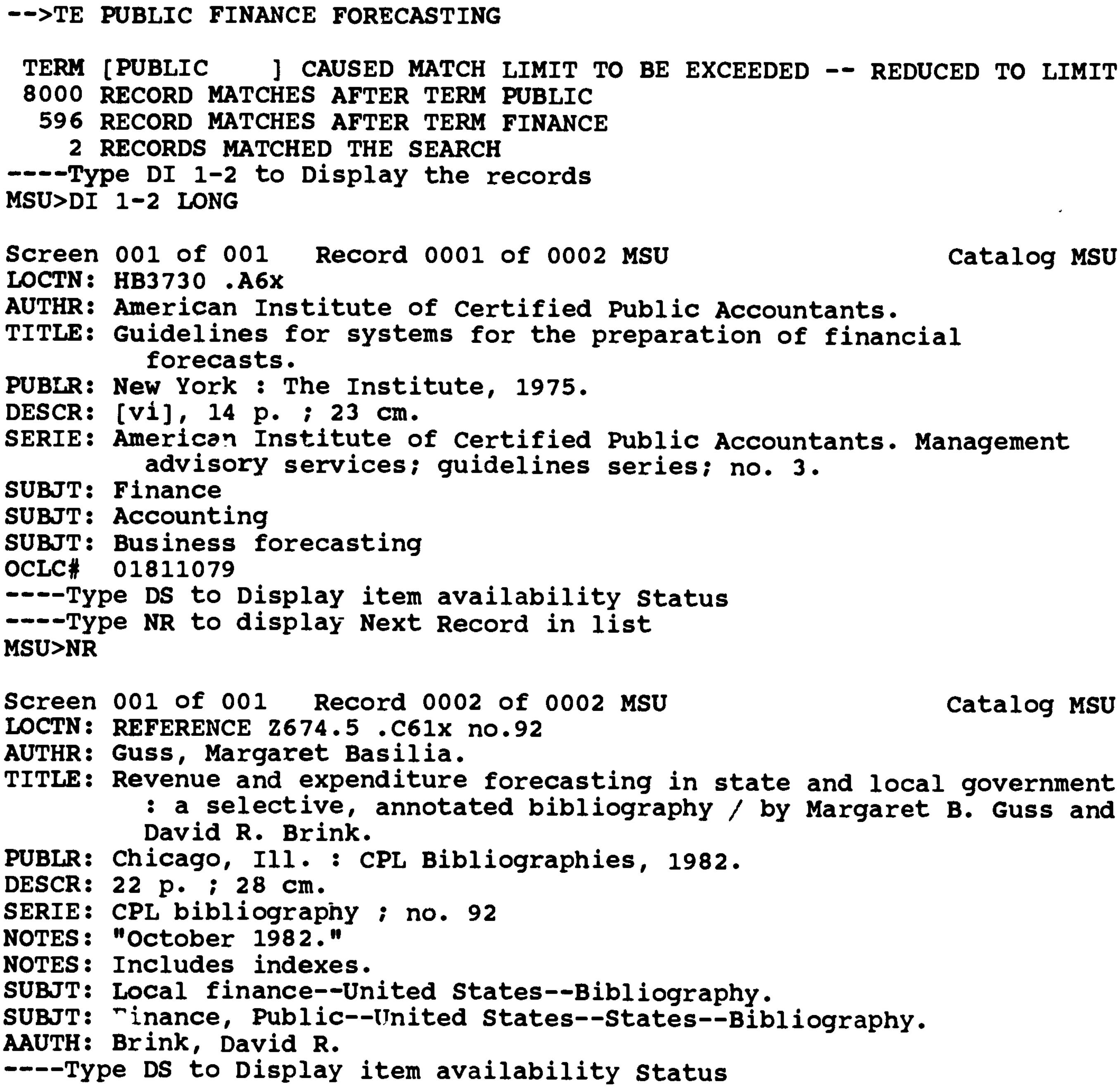
```
-->SU PUBLIC FINANCE

   NO RECORDS MATCHES THE SEARCH
 Try the TERM (TE) search -- it's more general
-->MSU>ST PUBLIC FINANCE

 3803 RECORD MATCHES AFTER TERM PUBLIC
   403 RECORDS MATCHED THE SEARCH
----Type DI 1-20 to Display first 20 records    (or)
Use AND command with additional WORD(s) or LIMITING command to reduce
results
MSU>DI 1 LONG

MESSAGE UNCLEAR - TRY AGAIN OR TYPE HELPER
>DI 1-3 LONG


Screen 001 of 001   Record 0001 of 0403 MSU                    Catalog MS
LOCTN: GOVERNMENT PUBLICATION C 3.145/4:977
AUTHR: United States. Bureau of the Census.
UTITL: Census of governments.
PUBLR: [Washington] : U.S. Dept. of Commerce, Bureau of the Census, [1978-
DESCR:    v. 28 cm.
NOTES: v. 1. Governmental organization.
SUBJT: Local officials and employees--United States.
SUBJT: State governments--Officials and employees.
SUBJT: Local finance--United States.
SUBJT: Finance, Public--United States--States.
SUBJT: United States--Administrative and political divisions.
OCLC#  03933486
----Type DS to Display item availability Status
----Type NR to display Next Record in list
```

**Figure 15.** *Subject Phase Search (SU) and Subject Keyword Search (ST)*
*(Note prompting for alternative search information.)*

a menu, or invoked by a pressed function key or simple typed command. Phrase searches are automatically processed as straightforward character string matches or word adjacency, same order matches. In most cases the match must begin with the leftmost significant word or character of the indexed entry. Keyword searches with more than one word are automatically processed as Boolean AND queries in most OPACs. Keyword searches may be targeted by the user or the system to one or more field indexes (e.g., title, notes, series statement, subject headings). Some flexibility is permitted in most keyword search OPACs. When searchers choose not to identify in their queries fields for subject keyword searching (or are not permitted by the system to do so), the trend is toward automatically searching both subject and title. ANDed keywords do not have to appear together in either the title field or the subject field to cause a match. Matching records would include those having one word in the title and another word in the subject heading.

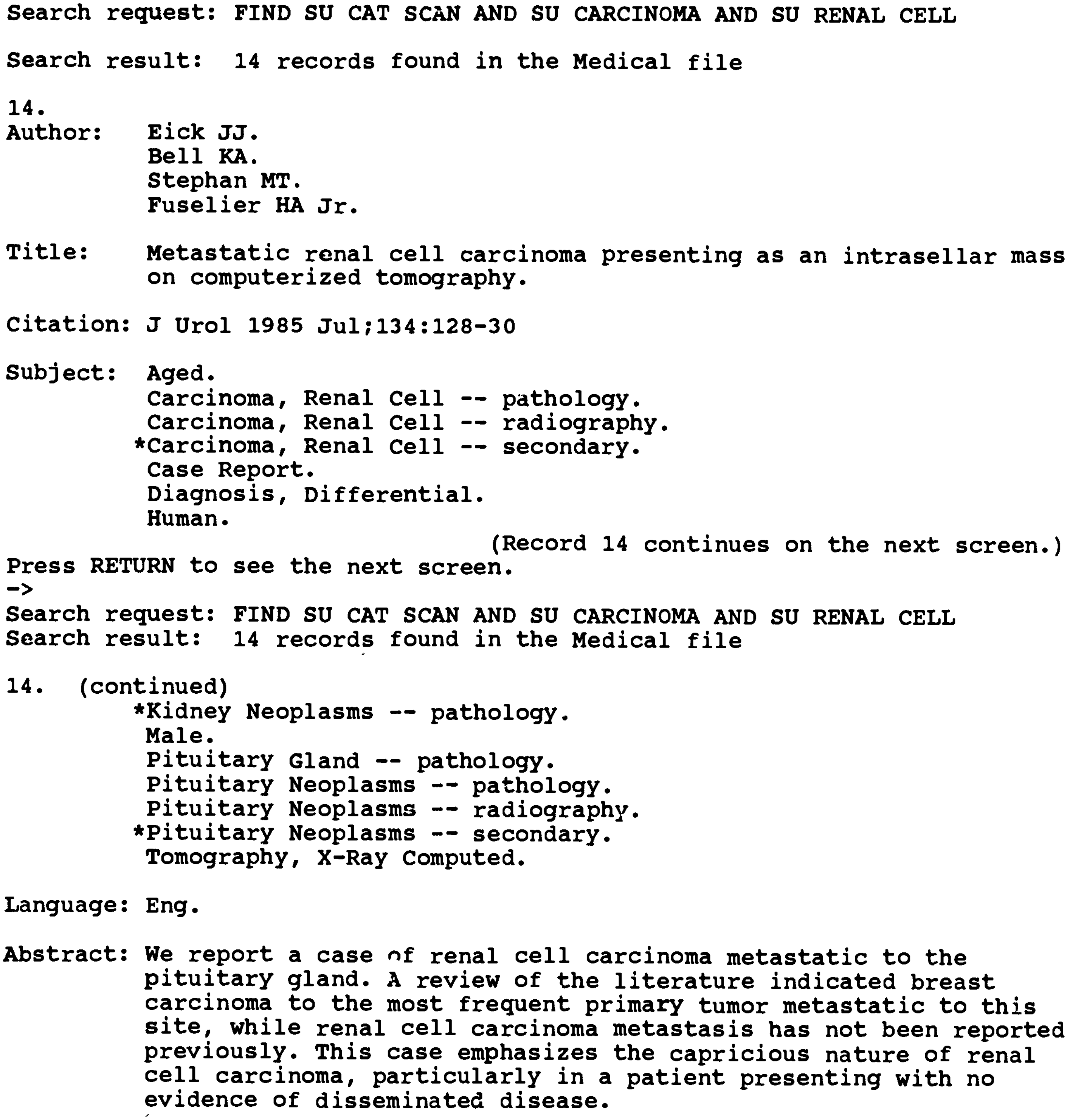
Successful subject phrase searching in most OPACs still requires an exact match on at least the initial, main portion of the subject heading in the catalog record. This usually requires a perfect match on a Library of Congress Subject Heading (LCSH). However, if a LCSH heading is entered as a search term and the heading has been modified or replaced in the local catalog record no match will occur. In some OPACs, when no match occurs on the user-entered term, the system displays headings in the alphabetical neighborhood of the term. This may or may not help the searcher find a heading which expresses his search interest.

This exact match requirement of OPAC phrase subject searches is illustrated in Figure 15 (above). The first search formulation, a subject phrase search, resulted in no matches because the main LCSH heading for this topic is inverted to "Finance, Public." When reformulated (as PALS suggests) as a subject keyword Boolean AND search (ST Public Finance), the query retrieved 403 citations. This illustrates the power of the keyword approach so useful when the exact form of the subject heading is unknown to the searcher. However, keyword searching has some drawbacks, especially when searches are applied broadly to many data fields in the bibliographic record. Figure 16 shows how broad (multiple field) subject keyword searches can produce "false drops" (retrieved bibliographic records not relevant to the user's topic of interest). Two records matched the query "exactly" as specified. The query merely required that each of the three individual words appear *anywhere* in the matching record. Thus, Record #1 is a legitimate match, but it is a false drop (also known as a "false coordination" or noisy match). Note also that records having two of the three words, for example, "public" and "finance," were not retrieved although they would probably be more relevant than Record #1. Boolean retrieval methods are not very intelligent. Figure 17 shows one way the keyword, Boolean approach can be aided by thesaurus-controlled searching and automatic cross referencing. "Cat Scan" is not in the bibliographic record, but it is linked to the preferred term in the Mesh (medical subject headings) Thesaurus.



```
-->TE PUBLIC FINANCE FORECASTING

 TERM [PUBLIC    ] CAUSED MATCH LIMIT TO BE EXCEEDED -- REDUCED TO LIMIT
 8000 RECORD MATCHES AFTER TERM PUBLIC
  596 RECORD MATCHES AFTER TERM FINANCE
    2 RECORDS MATCHED THE SEARCH
----Type DI 1-2 to Display the records
MSU>DI 1-2 LONG

Screen 001 of 001   Record 0001 of 0002 MSU                 Catalog MSU
LOCTN: HB3730 .A6x
AUTHR: American Institute of Certified Public Accountants.
TITLE: Guidelines for systems for the preparation of financial
         forecasts.
PUBLR: New York : The Institute, 1975.
DESCR: [vi], 14 p. ; 23 cm.
SERIE: American Institute of Certified Public Accountants. Management
         advisory services; guidelines series; no. 3.
SUBJT: Finance
SUBJT: Accounting
SUBJT: Business forecasting
OCLC#  01811079
----Type DS to Display item availability Status
----Type NR to display Next Record in list
MSU>NR

Screen 001 of 001   Record 0002 of 0002 MSU                 Catalog MSU
LOCTN: REFERENCE Z674.5 .C61x no.92
AUTHR: Guss, Margaret Basilia.
TITLE: Revenue and expenditure forecasting in state and local government
         : a selective, annotated bibliography / by Margaret B. Guss and
         David R. Brink.
PUBLR: Chicago, Ill. : CPL Bibliographies, 1982.
DESCR: 22 p. ; 28 cm.
SERIE: CPL bibliography ; no. 92
NOTES: "October 1982."
NOTES: Includes indexes.
SUBJT: Local finance--United States--Bibliography.
SUBJT: Finance, Public--United States--States--Bibliography.
AAUTH: Brink, David R.
----Type DS to Display item availability Status
```

**Figure 16.** *PALS General Keyword Search (TE)*



```
Search request: FIND SU CAT SCAN AND SU CARCINOMA AND SU RENAL CELL

Search result:  14 records found in the Medical file

14.
Author:   Eick JJ.
          Bell KA.
          Stephan MT.
          Fuselier HA Jr.

Title:    Metastatic renal cell carcinoma presenting as an intrasellar mass
          on computerized tomography.

Citation: J Urol 1985 Jul;134:128-30

Subject:  Aged.
          Carcinoma, Renal Cell -- pathology.
          Carcinoma, Renal Cell -- radiography.
         *Carcinoma, Renal Cell -- secondary.
          Case Report.
          Diagnosis, Differential.
          Human.
                                  (Record 14 continues on the next screen.)
Press RETURN to see the next screen.
->
Search request: FIND SU CAT SCAN AND SU CARCINOMA AND SU RENAL CELL
Search result:  14 records found in the Medical file

14.  (continued)
         *Kidney Neoplasms -- pathology.
          Male.
          Pituitary Gland -- pathology.
          Pituitary Neoplasms -- pathology.
          Pituitary Neoplasms -- radiography.
         *Pituitary Neoplasms -- secondary.
          Tomography, X-Ray Computed.

Language: Eng.

Abstract: We report a case of renal cell carcinoma metastatic to the
          pituitary gland. A review of the literature indicated breast
          carcinoma to the most frequent primary tumor metastatic to this
          site, while renal cell carcinoma metastasis has not been reported
          previously. This case emphasizes the capricious nature of renal
          cell carcinoma, particularly in a patient presenting with no
          evidence of disseminated disease.
```

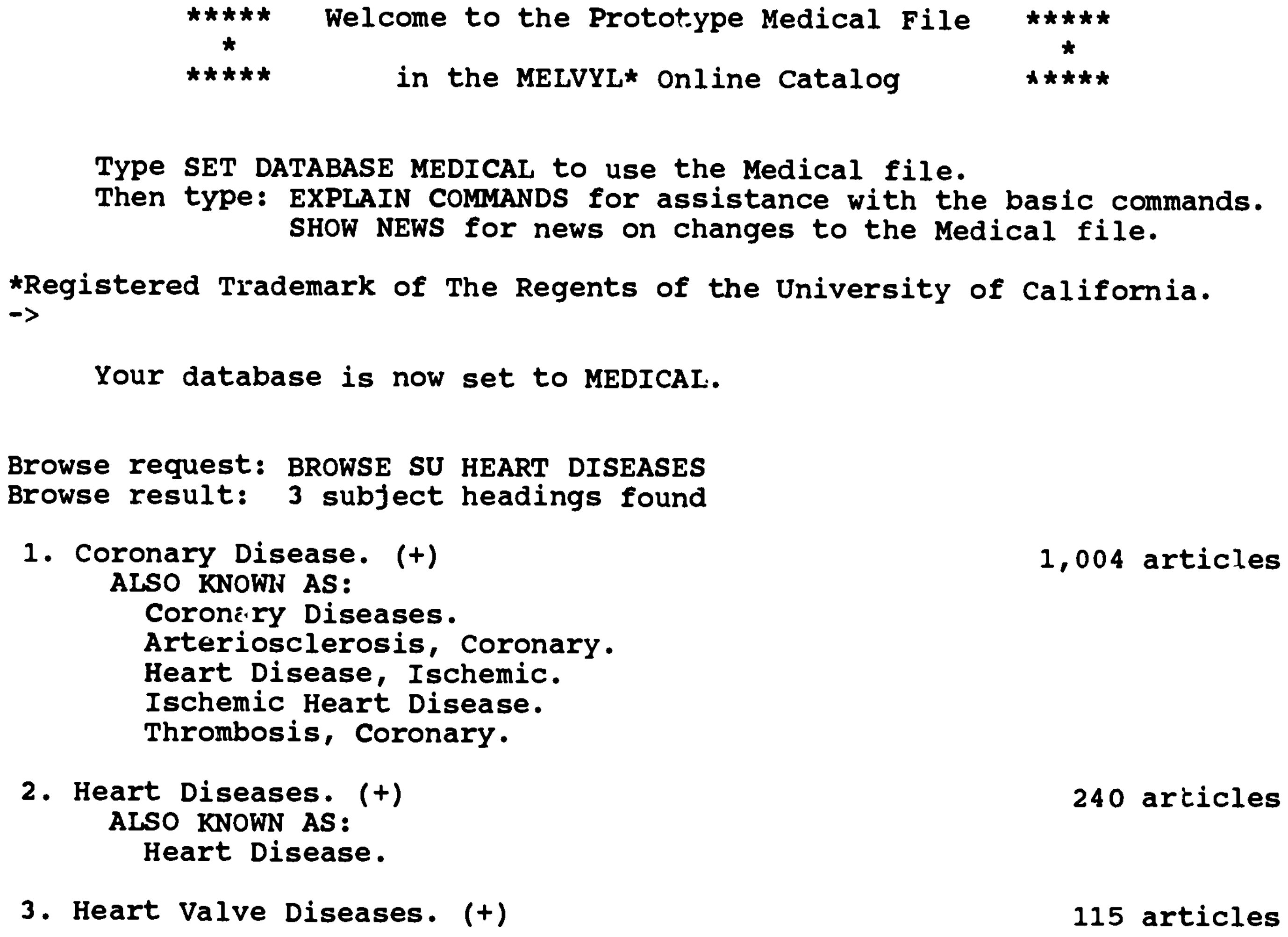
**Figure 17.** *MELVYL Search in MEDLINE*

In today's OPACs browsing is even less developed and is generally provided as an intermediate or secondary stage in the subject search process. Some keyword OPACs offer the searcher an unpublished (on the screen) option to browse (scan) alphabetical lists of index terms or headings. Phrase search OPACs typically display these lists *after* a search is constructed and entered, even when an exact match occurs. Research has not shown whether, in the latter case, this helps or confuses the searcher. A few OPACs are beginning to include cross references and related terms in these browsing displays (See Figure 18). This will make them more useful browsing tools for certain kinds of subject searching.

Figure 19 illustrates a shelf list browse display invoked by the entry of a truncated call number. The limited data presented in this typical array will not support the kind of actual shelf browsing practiced by library patrons in open stacks. Getting additional online information for a selected document, and then returning (if permitted) to the browse list can be cumbersome. There is no good reason to restrict shelf list browse entries to a single line of display data.

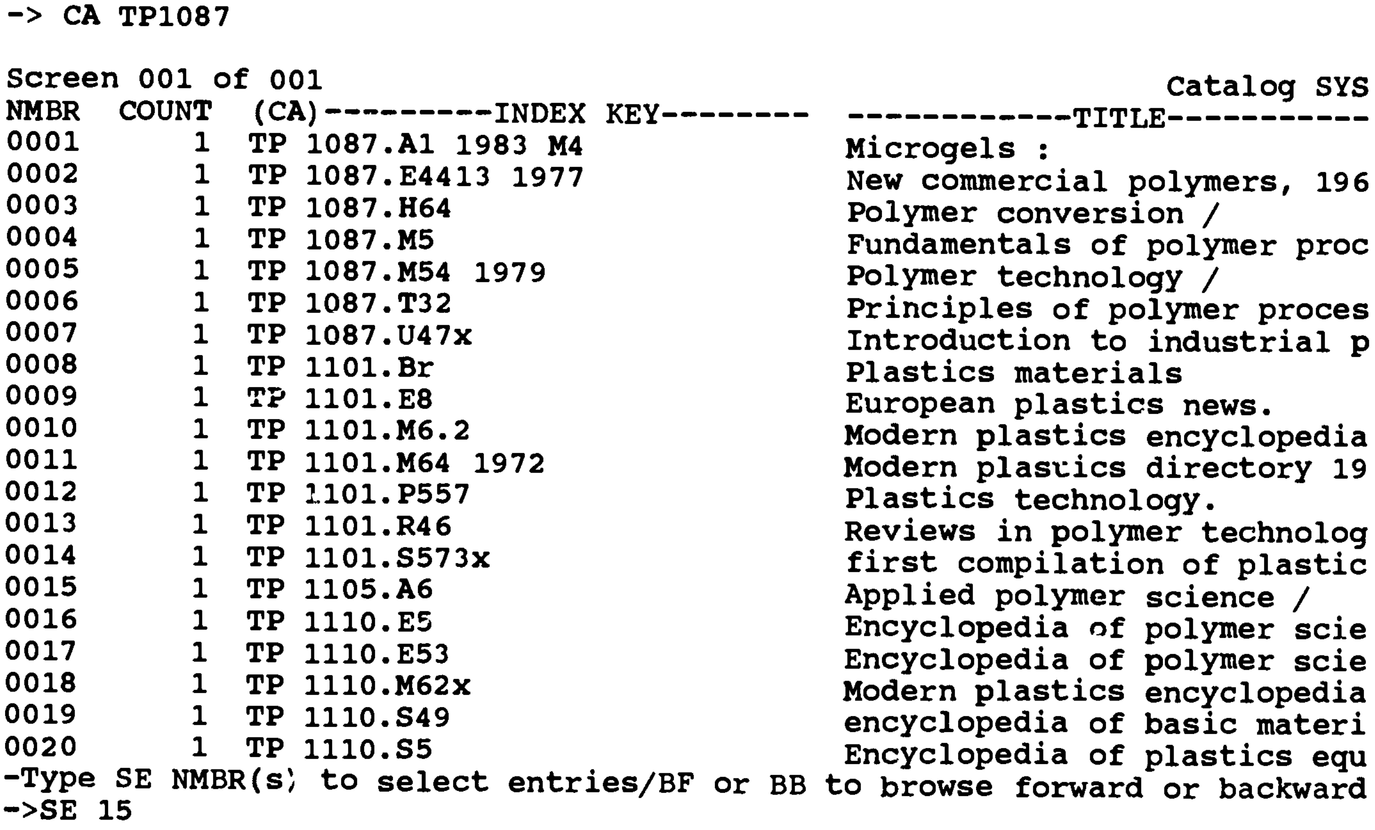
More informative, flexible, and intelligent browsing facilities can be provided in the online environment. It is time for librarians and OPAC designers to recognize that subject browsing may be a *primary* search activity in the user's quest to discover materials on a topic, or to discover unknown items of potential interest. This kind of search activity requires more than term selection browsing facilities, or even thesaurus-based automatic switching among related terms. Related *document* browsing and discovery can be facilitated in OPACs through richer precoordination in the database of multiple subject/topic clues found in bibliographic records (e.g., linking title terms with subject headings with call numbers, etc.), and by providing more search navigation options between retrieved and unretrieved (but linked) records, that is, record to record "jumping" at the discretion of the searcher (e.g., "Show me more books from this publisher." "What other titles are in this series?" "What documents cite this work?")

Today's OPACs also fail to provide more traditional subject search aids. For example, they do not support browsing in displays of classification outlines or schedules. Neither do they permit related term lookups in online subject thesauri or lists of subject headings. Markey's research has shown the first of these to be useful in OPAC subject searching.[10] Classification schedule terminology and systematic subject information can lead the searcher to relevant records not likely to be retrieved by the traditional phrase and keyword subject search methods. It is seldom recognized that even Cutter favored this approach. Immediately following the famous passage where he lists the "objects" of the library catalog, Cutter proposes the means for achieving them. To enable finding books on a subject, the catalog needs the subject entry, cross references, and a "classified subject table"![11]



```
          *****   Welcome to the Prototype Medical File    *****
            *                                                *
          *****        in the MELVYL* Online Catalog       *****


     Type SET DATABASE MEDICAL to use the Medical file.
     Then type: EXPLAIN COMMANDS for assistance with the basic commands.
                SHOW NEWS for news on changes to the Medical file.

*Registered Trademark of The Regents of the University of California.
->

     Your database is now set to MEDICAL.


Browse request: BROWSE SU HEART DISEASES
Browse result:  3 subject headings found

 1. Coronary Disease. (+)                                  1,004 articles
      ALSO KNOWN AS:
        Coronary Diseases.
        Arteriosclerosis, Coronary.
        Heart Disease, Ischemic.
        Ischemic Heart Disease.
        Thrombosis, Coronary.

 2. Heart Diseases. (+)                                      240 articles
      ALSO KNOWN AS:
        Heart Disease.

 3. Heart Valve Diseases. (+)                                115 articles
```

**Figure 18.** *MELVYL Browse in MeSH Thesaurus*



```
-> CA TP1087

Screen 001 of 001                                              Catalog SYS
NMBR  COUNT  (CA)---------INDEX KEY--------  ------------TITLE-----------
0001     1   TP 1087.A1 1983 M4               Microgels :
0002     1   TP 1087.E4413 1977               New commercial polymers, 196
0003     1   TP 1087.H64                      Polymer conversion /
0004     1   TP 1087.M5                       Fundamentals of polymer proc
0005     1   TP 1087.M54 1979                 Polymer technology /
0006     1   TP 1087.T32                      Principles of polymer proces
0007     1   TP 1087.U47x                     Introduction to industrial p
0008     1   TP 1101.Br                       Plastics materials
0009     1   TP 1101.E8                       European plastics news.
0010     1   TP 1101.M6.2                     Modern plastics encyclopedia
0011     1   TP 1101.M64 1972                 Modern plastics directory 19
0012     1   TP 1101.P557                     Plastics technology.
0013     1   TP 1101.R46                      Reviews in polymer technolog
0014     1   TP 1101.S573x                    first compilation of plastic
0015     1   TP 1105.A6                       Applied polymer science /
0016     1   TP 1110.E5                       Encyclopedia of polymer scie
0017     1   TP 1110.E53                      Encyclopedia of polymer scie
0018     1   TP 1110.M62x                     Modern plastics encyclopedia
0019     1   TP 1110.S49                      encyclopedia of basic materi
0020     1   TP 1110.S5                       Encyclopedia of plastics equ
-Type SE NMBR(s) to select entries/BF or BB to browse forward or backward
->SE 15
```

**Figure 19.** *PALS' Shelf List Browse Feature*

A few OPACs now display related terms associated with subject headings actually assigned to specific catalog records in the database. But no OPAC widely available provides access to an online subject thesaurus for concept relationship exploration or for the identification of broader and narrower search terms. Of course one explanation for this is the absence of a hierarchical structure in our major and most-used OPAC subject vocabulary, LCSH. An online catalog database that is linked to a hierarchically structured thesaurus - or any other systematic concept/term network structure - can support both far more sophisticated forms of browsing and far more refined, focused "known-subject" searching than is possible in existing OPACs.

## 3. Problems and Shortcomings of Today's OPACs

Good retrieval performance in second-generation online catalogs can be achieved only by library staff and by library patrons trained to use and understand their particular indexing and search idiosyncracies. Most of these online catalogs are not yet effective, usable "self-service" information retrieval systems for a wide variety of untrained, occasional users.

Bates poses the central question about subject access in today's online catalogs: "With all the power of online subject searching of catalogs - Boolean logic, keyword match, truncation, etc. - have we, perhaps, already given the user all the search capability that is practically necessary?"[12]

Online catalog research studies have uncovered a number of common problems experienced by users of second-generation online catalogs. In general terms the major problems include:

- too many failed searches (search attempts that are aborted, that result in no matches, or that result in unmanageably large numbers of items retrieved),[13]

- navigational confusion and frustration for the user during the search process ("Where am I?", "What can I do now?", "How can I start over?"),[14]

- unfamiliarity with subject indexing policy and vocabulary, leading too often to the failure to match search terms with the system's subject vocabulary,[15]

- misunderstanding and confusion about the fundamentally different approaches to retrieval and search methods employed in today's online catalogs (e.g., precoordinate phrase searching and browsing, and postcoordinate keyword/Boolean searching),[16] and,

- partially implemented search strategies and missed opportunities to retrieve relevant materials (e.g., searches in which large retrieval sets are not scanned or narrowed in size, and title keyword searches that are not followed by searches on the call numbers or subject headings of the found records).

Chan points out that online searching is a process of extracting a subfile of useful documents from a large file, a process where "in most cases, a sequence of search statements is required for even minimally satisfactory retrieval."[17] To optimize retrieval results in subject searching, more than one search approach may have to be employed in the overall search strategy: "Through combination, keywords and the [controlled] vocabularies of DDC, LCC, and LCSH should offer far greater possibilities in search strategies than any one of them can provide alone."[18] Markey has demonstrated, for example, that different records on a particular subject would be retrieved by using a classified approach from those retrieved using keyword or alphabetical subject heading browsing approaches.[19]

## PROBLEMS AND SHORTCOMINGS

Conventional information retrieval systems place the burden on the user to reformulate and reenter queries until satisfactory results are obtained. This is typically the case with second-generation online catalogs, as well. This approach assumes, however, that the user *knows* what he wants and can describe it in the language of the catalog database being searched.

Even the best second-generation catalogs do little to help the user transform an information need to explicit expressions of the need acceptable by the system. Nor do these catalogs lead the user from "found" information to related, linked information that has not yet been discovered. It is unrealistic to expect our catalog users to know in advance the structure and language of our library databases. It is equally unrealistic to expect online catalog users to be proficient in the various search approaches and techniques *before* they engage an interactive system in the retrieval process. Humans find it easier to recognize things than to generate formal descriptions. OPACs could take advantage of this human facility by permitting requests such as, "Give me more like this!"

### LCSH-Based Subject Searching

Markey's research has shown that online searchers are not very successful in matching their subject terms with the catalog's controlled subject vocabulary.[20] The assigned subject headings in the catalogs she investigated were derived from the list of Library of Congress Subject Headings (LCSH), the subject vocabulary used in most library catalogs. In one study of subject searching on a university OPAC (SULIRS, Syracuse University), a total of 859 search statements entered in 188 subject searches were analyzed. 45% of these search statements resulted in no retrievals. "We were concerned about, and subsequently sought an explanation for, the surprisingly large percentage of subject searches resulting in no or very few retrievals." In comparing the subject statements with LCSH headings, only 29% of searchers' terms matched or closely matched LCSH. Yet only 7% of these LCSH searches resulted in no matches. This is not surprising since OPAC subject headings are extracted from actual bibliographic records in the collection. More often than not searchers used whatever terms popped into their minds; this produced no retrievals *65%* of the time. These depressing findings have been corroborated in studies of other university OPACs.

Another finding in Markey's SULIRS study raises questions about the effectiveness of LCSH subject searches for retrieving only the most relevant documents for a query. 29% of the LCSH searches retrieved 100 or more citations. When this occurred, searchers either aborted their searching or looked individually at all the results. Very few searchers entered valid LCSH cross references. A good guess is that they usually entered and matched LCSH terms broader than their topic of interest. Markey concludes that the study "drives home the need for online vocabulary assistance. Alphabetically arranged or rotated lists of subject headings only scratches the surface."

Based on a variety of OPAC subject searching research studies, Markey lists the major problems subject searchers face, with little or no assistance from OPACs:

- matching their terms with those indexed in the catalog,
- identifying terms that are broader or narrower than their topic of interest,
- improving search results when little or nothing is retrieved,
- reducing search results when too much is retrieved,
- grasping LCSH indexing policies and idiosyncracies.

Pritchard underscores the "conceptual problems" facing users of the Library of Congress' online subject access systems (LOCIS): "To choose the best [LCSH] headings, readers may need to understand concepts of indexing depth and specificity since LC only catalogs for the general subject of the entire work and does not assign books to both broad and narrow categories. This causes difficulty both in topical headings and geographical ones. LOCIS searchers intuitively seek to post-coordinate numerous terms, whereas MARC records contain two or three (at best) pre-coordinated, subdivided, inverted, and parenthetically qualified headings."[21]

Bates' research also indicates that there is more to good subject searching than matching an LCSH heading. One study "tested whether people would actually hit upon specific relevant material in a search rather than just whether their term would match with any heading in the catalog, whether or not that heading indexed material relevant to their query."[22] Participants in the study were asked to state what word or phrase they would use to search in the subject catalog for a specific book described to them by its title and an abstract. The degree of match between their terms and the subject headings used in the library catalog to index the books was calculated. The initial set of contributed terms matched headings in the catalog record only slightly more than 20% of the cases. Thus, there are two (at least) major online subject access problems associated with the official subject vocabulary (LCSH): 1) initially matching the assigned headings (the "entry" or "lead-in" vocabulary problem), and 2) once into the system's vocabulary, identifying terms and term relationships that will help direct a search to the precise topic or most relevant materials (the subject focusing/discriminating problem). Mandel and Herschman point out that OPAC subject browsing displays that list only subject terms from catalog records in the system do not show conceptual relationships among terms or lead the searcher to the most relevant term.[23] A linear alphabetical arrangement of terms scatters related terms. The authors recommend including a hierarchical subject thesaurus online - a restructured LCSH - to help searchers overcome this problem.

## PROBLEMS AND SHORTCOMINGS

With regard to the entry term, LCSH matching problem, Mandel and Herschman explain: "Displaying the existing LCSH headings and references will not, in itself, solve the problem of an entry vocabulary that matches only half of users' first tries. Even without making a single change in an LC subject term, access to the terms could be improved enormously by adding to the entry vocabulary (i.e., adding "see" references).[24]

Bates describes this as the problem of low vocabulary "redundancy" (i.e., lack of quantity and variety of synonyms and related terms) in subject catalog records stemming directly from LC's subject indexing policies and practices.[25] These book indexing/subject cataloging policies - precoordinate, whole book indexing; single, uniform heading indexing; specific heading only indexing - result in a very small amount of subject terminology in catalog records (on average, two or fewer headings per record). Indexing a particular book under broader or narrower terms is rarely done. As Bates says, "the searcher is only directed to terms at the same or more specific levels. Usually, there are not very many of even these latter references anyway."

Research has shown that subject searchers frequently enter terms broader or narrower than the subject in which they were actually interested. LCSH offers little assistance here in its present form, confirmed by the low LCSH term match ratio in OPAC subject searching.

The LCSH cross reference structure is limited and weak. Cochrane believes that following its "see also" references usually leads the searcher out of his topic, rather than to various aspects or levels of the topic.[26] Bates distills an expression of the dilemmas experienced by subject searchers in this poignant passage: "In both manual and online catalogs the user must launch the search with a subject term. For the search to be successful, the term must not only match with some term in the system, but must match, either directly or through cross references, with a term describing *relevant* material. As I have argued earlier, LCSH uses so few terms for indexing each document and provides so little assistance to the searcher that the latter is hard to do."[27]

In summary, second-generation online catalogs are subject-searching weak because they:

** do not sufficiently assist with the translation of entered query terms into the vocabulary used in the catalog,

** do not provide online thesaurus aids useful for subject focusing and topic/treatment discrimination,

** do not automatically assist the user with alternative formulations of the search statement or execute alternative search methods when the initial approach fails,

** do not lead the searcher from successful free-text search terms (e.g., title words) to the corresponding subject headings or class numbers assigned to a broader range of related materials,

** do not provide sufficient information in the retrieved bibliographic records (such as tables of contents, abstracts, and book reviews) to enable the user to judge the usefulness of the documents,

** do not rank the citations in large retrieval sets in decreasing order cf probable relevance or "closeness" to the user's search criteria,

** do not facilitate open-ended, exploratory browsing through following pre-established trails and linkages between records in the database, in order to retrieve materials related to those alre1dy found.

### Boolean Retrieval Methods: A Reappraisal

In the early 1980s many librarians and OPAC researchers urged vendors and system designers to upgrade their OPACs to keyword, Boolean retrieval systems. Many thought that the provision of Boolean search formulation and retrieval methods in OPACs would provide more search flexibility and subject access points than first-generation OPACs, and, by offering term/quer/ postcoordinated searching, would give subject searchers an effective alternative to exact matches on unknown LC subject headings. Some librarians welcomed keyword/Boolean OPACs as the panacea for the problems of subject searching in early OPACs.

This enthusiastic anticipation of the arrival of Boolean OPACs in libraries is easy to explain. First-generation OPACs were not very good retrieval systems. Many of them are still in place alongside the new interactive CD-ROM reference retrieval systems and the online search terminals being used to access the commercial database search services. Boolean retrieval is the predominant mode of access in the world of commercial online reference searching, and has acquired over the years an i.nmense prestige in the eyes of librarians. The commercial database search services have steadily grown in the number and size of their databases, and ihe number of libraries and librarians using one or more of them has increased greatly in the 1980s. Doszkocs comments on this phenomenon:

> The impressive growth and acceptance of these systems is partly due to the search efficiencies inherent in the inverted list file structures and Boolean set operatio:ns commonly employed. The prime advantages of the inverted file, Boolean logic design paradigm are speed

> and iterative search flexibility. These advantages, however, are invariably offset by limitations in query and document analysis and the restrictive nature of the user interface. The file inversion process inevitably results in a certain loss or increased ambiguity of meaning in searching document content, regardless of whether manual or automatic indexing procedures are utilized. Similarly, the Boolean coordination of search terms imposes semantic loss and a yes-no rigidity on matching queries to document ti.les, abstracts, full text or subject descriptors.[28]

Reflecting on the popularity of Boolean retrieval, Porter and Galpin remark, "This is unfortunate, since it has a number of inherent weaknesses."[29] What is behind this prevalent anti-Boolean opinion which seems to pervade the IR research community? In a nutshell it is this: much research and experience with Boolean retrieval systems (including OPACs) indicates clearly and repeatedly that Boolean search formulation syntax and retrieval techniques are not very effective in search performance and not very usable or efficient search methods for end users. The accumulating evidence clearly supports this summary critique of Boolean retrieval by Porter and Galpin:

> The number of documents retrieved is usually too large or too small, and a certain amount of juggling with terms is necessary to get a retrieved set of manageable size. Users frequently cannot compose boolean expressions, and require an expert to do it for them. The retrieved set of documents is usually not ranked in any way, so it is necessary to inspect the entire list in the search for relevance.[30]

In his recent review of the literature on end-user searching of online bibliographic databases, Mischo reports: "There is also much anecdotal evidence and observation showing that end users have particular difficulties with search-strategy formulation and the use of Boolean logic."[31] Mischo discovered that numerous authors tell of the significant difficulties end users are experiencing with the proper use of Boolean logic: "Several evaluation studies indicate that the use of Boolean operators is viewed as the most difficult aspect of retrieval."[32] Apparently, the solution is not more and better training programs. (How do you force training on public library patrons and dial-up OPAC users?) In one study reviewed by Mischo, the experimental end-user group was required to attend several training sessions and have their search strategy approved prior to searching online (BRS). The investigators found that end users had major problems with the choice of terminology, the use of Boolean operators, and search strategy modification. Even after this training and supervision, each subject who returned to attempt new searches needed to meet with expert search specialists beforehand to review system commands and Boolean logic.[33] These findings have been corroborated by a number of similar studies.

In explaining their motives in designing an alternative, non-Boolean, natural language retrieval system for their users ("STATUS with IQ"), Pape and Jones refer to the "basic problem" with Boolean logic systems: "namely that high precision and high recall seem incompatible in this environment." And they go on: "There is also the important issue of allowing queries to be entered in natural language and saving users from the horrors of a typical boolean based query syntax."[34] Salton, *et al*, describe the high recall/high precision compatibilty problem in this way:

> The basic problem in Boolean query formulation consists in first choosing an appropriate set of query terms, and in then using the Boolean operators to generate a formulation which is not so broad as to retrieve an unreasonable amount of extraneous matter thereby causing a loss in search precision, nor so narrow as to reject a large number of relevant items thereby causing a loss in search recall.[35]

Noreault, *et al*, the founders of SIRE, the prototype for OCLC's new CD-ROM bibliographic retrieval system, focus their criticism of traditional Boolean retrieval on a major problem associated with the output (results) of Boolean systems. The problem is especially felt by users when their searches produce large results sets:

> In a normal Boolean search output one expects a random distribution of relevant documents. That is, Boolean searching of a data base usually results in a list of references with no indication as to which of those documents are more likely to be relevant to the user's request. Ranked output [on the other hand] attempts to provide the user with information indicating that the closer a document is to the beginning of the output list, the more likely it is to be relevant to his query.
>
> Automatic ranked output based on probable relevance requires the calculation of a similarity measure which does more than make a binary decision on whether there is any matching of terms between the query and the document.[36]

A near-consensus exists among information retrieval theorists and investigators regarding the shortcomings of IR systems that rely solely on Boolean-logic query formulation and matching. Salton gives us the best overall summary of criticisms of Boolean retrieval systems voiced or shared by most researchers and experimenters in the information science community:

1. The formulation of good Boolean queries is an art rather than a science; most untrained users are unable to generate effective query statements without assistance from trained searchers.

2. The standard Boolean retrieval methodology does not provide any direct control over the size of the output; some query statements may provide no output at all, whereas other statements provide an unmanageably large number of retrieved items.

3. The Boolean methodology does not provide a ranking of the retrieved items in any order of presumed usefulness, thus all retrieved items are presumed to be equally good, or equally poor, for the user.

4. The Boolean system does not provide for the assignment of weights to the terms attached to documents or queries; thus each assigned term is assumed to be as important as each other assigned term, the only distinction actually made is between terms that are assigned (with an implied weight equal to 1), and terms that are not assigned (with an implied weight equal to 0).

5. The standard retrieval methodology may produce results which appear to be counter-intuitive:

   a. in response to an or-query (A or B or ... or Z) a record or document with only one query term is assumed to be as important as a document containing all query terms;

   b. in response to an and-query (A and B and ... and Z) a document containing all but one of the query terms is considered as useless as a document with no query term at all.[37]

Online catalog research and design has been directed to making post-coordinate, Boolean logic, library retrieval systems easier to learn and easier to use than the commercial models used by trained intermediaries. However, little attention has been given to the major performance limitations of Boolean OPACs.

## 4. Information Retrieval Research and Experimentation: Perspectives, Insights, and Contributions

Over the past fifteen to twenty years researchers in automated information retrieval have contributed a large body of experimental findings, theory, and published literature.[38] There have been many advances in the field, and the work of the IR researchers has produced a great many successful experiments and enlightening results. The primary accomplishment of this group of researchers and scholars has been the transformation of traditional indexing and retrieval analysis, opinion, and design activities into an empirical science resting on sound theoretical bases. Major outcomes of this scientific work include: 1) the development of a deeper understanding of the inherent complexities in the information retrieval process and surrounding situation, 2) new theoretical models of the IR environment, models which have more explanatory and predictive power, 3) widely applicable evaluation methods and performance measures, and 4) tested, more effective retrieval techniques and more usable user-system interfaces.

### The Information Retrieval Situation

Eastman provides a useful description of the information retrieval function and its constituent processes as understood by information scientists:

> Although the information retrieval architecture has been used in a variety of contexts, the archetypical system is one designed to handle document retrieval. In response to user queries, the system retrieves documents relevant to those queries. So the queries correspond to problem instances, and the documents correspond to possible solutions. A common representation for both queries and documents is as sets of keywords, or index terms. A query is abstracted into a set of keywords to be used as search terms. It is then matched against document representations to choose documents that appear likely to be relevant. Most current commercial systems handle queries that are represented as Boolean combinations of keywords. A number of experimental systems are based upon alternative representations, including vector space representations, and use different matching algorithms.
>
> Heuristic searching is almost always present, but may be shared in a variety of ways between the system and the searcher, who is frequently an intermediary between the user and the system. The search query may be expanded by considering broader terms (super classes), narrower terms (subclasses), or related terms (synonyms or siblings). This expansion may be done by using a thesaurus or by examining intermediate output. The search is generally performed interactively and modified on the basis of intermediate results. In commercial systems, the searcher is responsible for most of the heuristic searching. However, ways to handle it automatically are being investigated.[39]

## INFORMATION RETRIEVAL RESEARCH

There is wide agreement among information scientists that the Boolean retrieval model is theoretically flawed because it does not reflect or account for all the inherent subtleties and complexities which comprise the real world information retrieval situation. Researchers have proposed alternative models of the IR function and situation which they believe more accurately identify critical aspects of the problem area under study. As Bookstein explains, "One of the most important functions of a model, mathematical or otherwise, is that it helps us focus our attention on features of a problem area that may have been overlooked when simpler models are considered. It gives us a way of thinking about the problem."[40] IR researchers believe that research-based design improvements in these critical areas will lead more often in practice to the attainment of the fundamental goal of IR system use. (See Figure 20).

USER

↓

INFORMATION NEED/PROBLEM

NEED/PROBLEM NEGOTIATION, CLARIFICATION, EXPRESSION

QUERY TERM SPECIFICATION/SELECTION

QUERY STATEMENT FORMULATION/MODIFICATION

QUERY PROCESSING

RESULTS REPORT/DOCUMENT SURROGATES DISPLAY

RETRIEVAL RESULTS REVIEW AND RELEVANCE ASSESSMENT

↑

DOCUMENTS

**Figure 20.** *The Information Retrieval Situation*

(*Note*: IR System Aim - to identify documents or other sources of information likely to be relevant and useful, with regard to a user's information need or problem, for assessment and possible use by the user.)

If theory is to lead to improvements in practice (i.e., IR system design and use), theoretical models must take into account both the simple and the complex characteristics of the activity being modeled. Much of the complexity of the IR situation can be attributed to the large degree of indeterminacy, uncertainty, and variability inherent to various levels of the whole domain.[41]

Researchers have shown that the IR situation is loaded with variability at all sides and, as a result, uncertainty must be accepted as intrinsic to the retrieval process. From document description and subject analysis of texts to IR system design, efforts to improve matters must confront the inherently probabilistic nature of the entire retrieval environment. The problem is complex and has many dimensions. No single "solution" is waiting to be discovered, even with the coming of the intelligent interface.

Doszkocs describes the challenges facing IR researchers and system designers: "Investigators have been confronted with the variability of ways in which the same ideas and topics can be expressed by different authors, abstractors, indexers, and searchers, the inevitable limitations of the query- matching procedures and the contextual subjectivity of users' relevance judgments concerning retrieved items." Doszkocs characterizes the common goal of most IR researchers: "to transcend the limitations of the basic keyword/subject heading/inverted file/Boolean logic search paradigm characteristic of the mechanized systems of the 1960's and 1970's." In the process, "IR researchers have come to recognize the inherently uncertain and probabilistic nature of the information retrieval process."[42]

Bookstein further clarifies the impact of this pervasive variability and uncertainty on the fundamental task of information retrieval, namely, how to decide, on the basis of a variety of imperfect indicators ("clues") of document relevance, which documents to retrieve and the order in which to display them. The task is seen as a problem amenable to statistical decision theory solutions:

> Uncertainty seems to be a characteristic intrinsic to the information retrieval (IR) process. When retrieving a set of documents in response to a request, an Information Retrieval System (IRS) must somehow make decisions about document relevance on the basis of items of evidence, each of which only imperfectly indicates the appropriate action to take. A range of data is currently recognized as being valuable for this purpose; these include such variables as author's name and institution, journal, age of paper, cited or citing papers, and, of course, indicators of subject content. However, even in situations in which only some or even none of these data alone provide strong evidence, collectively they may produce a rather strong case for retrieving or not retrieving a document. The problem, then, is how to bring together a range of evidence to make a retrieval decision in the face of uncertainty.[43]

## INFORMATION RETRIEVAL RESEARCH

The challenge facing system designers is to exploit the science and technology (affordable) of automated information retrieval to achieve the "best" retrieval for a given user query in an inherently imprecise and uncertain situation. Compounding the variabilities and complexities of subject cataloging/indexing, file structure, and matching and retrieval algorithms, the user may not know or be able to adequately express his need, or may simply change his mind during the retrieval process about what he wants or is interested in. In addressing the topic, "What is intelligent information retrieval?", Croft acknowledges the many advances made in the field of information retrieval since the arrival of the computer, but points to several basic issues remaining to be resolved. "To put it simply, we do not know the best way of representing the content of text documents and the user's information needs so that they can be compared and the relevant documents retrieved."[44] Croft points to the small but significant improvements to the retrieval process where statistical approaches to the analysis of text and collections of documents have been applied.

### Information Retrieval as an Inference Process: From Matching to Relevance

Subject access researchers like Bates and Markey have identified the major shortcomings of systems which do no more than execute exact matches of phrases or queries expressed as Boolean combinations of keywords and retrieve documents that contain exactly the phrase or combination of keywords entered. Simply put, such systems do not go far enough. The aim of any retrieval process is to bring relevant documents to the searcher, arranged in some useful manner so that they can be assessed. Conventional retrieval system matching mechanisms which exploit the inverted file structures of their databases may be internally efficient, but they too often produce large, unordered results sets that turn users off and away. Few end users display the desire or ability to use existing system query syntax to modify their queries to achieve better, more manageable results.

For these and related reasons researchers argue that information retrieval should be viewed as much more than an efficient, fast document matching and gathering operation. The IR situation requires that we view information retrieval as an iterative, truly interactive, inductive process, a process which engages the user throughout the process to gain relevance feedback that can be used by the system to correct its assumptions or to modify its automatically applied, heuristics-based matching and document ranking procedures. In other words, information retrieval, especially document retrieval, should be viewed as an interactive, cooperative process of mutually supportive inference.

Croft and Thompson draw a useful contrast between document or bibliographic retrieval systems and the database management systems now so popular in microcomputer software business applications. The retrieval facilities are similar, Croft points out, but document retrieval should not be viewed as a special case of data retrieval from such database systems. To do so "obscures the features of document retrieval that make it a challenging and difficult research area."[45]

What are these unique and troublesome characteristics of the document retrieval situation which force upon us the informed view that this sort of information retrieval is a process of inference? Users have a wide range of both predictable and unpredictable information needs. Most bibliographic, document searching appears not to be for previously known, specific items. Only in a small proportion of searches are users able to provide a query that accurately expresses their information needs.

Croft and Thompson explain that there is a big difference between a database query such as:

Find all employees with age >30 and salary <20000

and a bibliographic retrieval query such as:

Find all documents about controlling inflation through monetary policy

Far from providing an exact specification of the desired citations, the bibliographic retrieval query provides "only an indication of the content of the desired document. The actual content of the documents identified as relevant by the user may vary considerably from the phrases provided in the query."[46]

The aim of a bibliographic retrieval system is to retrieve documents likely to be relevant to a particular user's query, or, more precisely, documents relevant *in the eyes of the user.* Thus, relevance is a function of user assessment and cannot be established by the simple, mechanistic query- document matching procedures employed in conventional retrieval systems. We do not know enough about how and on what basis users make relevancy/utility judgments about retrieved bibliographic citations. Such reasoning activity may consist of a careful process of inference after examination of all the pertinent data in a citation. On the other hand it may consist of simple "flash" recognition, a drawing on a quick analogy to other known items, or just playing a hunch. The first case seems to be ruled out in present-day OPAC subject searching: every study of transaction logs indicates OPAC searchers seldom if ever look at a display of the full citation, which is the only way to find any *explicit* relevancy assessment data other than that contained in the title. The less subject data there is in the citation, the less likely a systematic process of inference will be undertaken to decide the matter. The user's knowledge of the subject field and any prior knowledge of the contents of the database would no doubt be significant variables in this assessment/selection activity.

An information retrieval system is effective to the degree that it supports and facilitates these document-relevancy assessment, selection or rejection activities. Since this human reasoning/recognition activity is not singular, one-dimensional, or usually predictable in a mechanistic way, it is unlikely a matching mechanism that does not interactively seek clues and rank its output will get the job done.

A number of methods have been tested for supporting this inference process, including automatic indexing techniques and retrieval techniques that employ statistical criteria and procedures. Statistical properties of text or terms in a database of citations are used to assign special values or weights to words, phrases, groups of related words, or clusters of citations. These techniques are in turn used in probabilistic or extended Boolean retrieval methods. The probabilistic model of retrieval, simply understood, infers the probability of relevance of documents in a given collection to a specific query and ranks them accordingly. The ranking algorithm orders the set of retrieved documents according to their decreasing similarity to the query. The probability of relevance may be calculated from the frequencies of occurrence of query/index terms in the entire database and/or the retrieved documents, or on the basis of a variety of other query--document similarity measures. As an example of term weighting, a term that occurs with very low frequency in the entire database but has a high occurrence count in particular documents would be considered to have special (high) value, and the documents it indexes would be considered to have a high probability of relevance.

Relevance feedback from the searcher is now considered essential to the effective performance of probabilistic retrieval systems. The searcher may explicitly change the values (system calculated weights) assigned to search terms or may respond to the first-listed, top-ranked documents. Relevance feedback may lead to a refinement or expansion of the user's query and "fuel" the system for even better performance.[47]

Probabilistic retrieval with relevance feedback is especially useful and effective in searching bibliographic databases because the user, on his own, cannot possibly know or specify all the possible linkages, associations, and relevancy ties among documents in a large multidisciplinary database. Probabilistic retrieval techniques, automatic search heuristics, and relevance feedback can exploit precoordinated conceptual structures and statistical associations to improve retrieval in such a universe.

Croft and Thompson summarize the advantages of probabilistic, statistical retrieval techniques:

* They are efficient to implement,

* They are more effective in terms of finding relevant documents than searches based on Boolean queries/exact matching,

* They have a sound theoretical basis,

* They are independent of any particular domain. That is, different types of documents (journal articles, office memos) from different domains (medicine, law) can be handled using the same techniques.[48]

Probabilistic, combinatoric, retrieval methods, and rule-based search strategy selection (if one retrieval strategy fails, automatically attempt another) can supplement the human tasks of relevancy assessment, inference, and selection better than Boolean methods, but neither can replace the human factor entirely. Human judgement is not only richer, it is the human who wants the documents or the information they contain. An intelligent retrieval system may never have the proper motivation to do a perfect job, that is, retrieve all relevant documents (assuming a comprehensive search is desired) and no non- relevant documents and rank order the retrieved documents according to degree of relevance. Croft and Thompson remind us that the other source of imperfection in any machine retrieval environment is the system's inability to achieve in its interpretation of a query anything more than a close approximation to the actual information need. He concludes that the two best ways to improve retrieval performance are "to enhance the inference process used by the system and to acquire better descriptions of the information need."[49]

## Summary of Contributions From Information Research and Experimentation

IR research has resulted in a significant gain in our knowledge of the information retrieval process and environment, more effective and feasible retrieval methods, and useful performance evaluation measures and methods. One of the strengths of the IR research tradition is the community's emphasis on testing and evaluation. New techniques and hunches have been put to the test, and have often produced negative results. Theories and new models are not accepted until there is a large body of confirming evidence. The two most common measures used to evaluate retrieval system effectiveness and performance are recall and precision. They are used to assess the results of specific retrieval operations. *Recall* is the proportion of relevant documents in the database retrieved; *precision* is the proportion of retrieved documents that are relevant.

These measures have been subject to some criticism that is often beside the point. Clearly there is an element of subjectivity in the relevancy judgments they measure. But once those judgments are made, recall and precision are quantitative measures that can be objectively applied in a given case. They are *standard* measures used to compare automated indexing techniques, retrieval methods, or systems in test or evaluation scenarios. They are not intended to replace or assume human judgement in real world retrieval environments. When applied consistently, especially to evaluate different retrieval strategies or methods in the same test document database, they are solid indicators of performance levels and they can support sound judgments regarding the relative effectiveness of one approach over another. They can also be used to measure incremental improvements resulting from refinements in a single technique or approach.

Doszkocs cites the following advanced information retrieval functions and features as being among the paramount achievements of the IR research and experimentation community: "the notion of accepting unrestricted natural-language user queries, flexible matching functions, ranking of retrieval output according to potential relevance to the query, and dynamic utilization of user feedback in automatic search strategy modification."[50]

Some newer IR systems and OPACs and demonstration prototypes have established the operational feasibility of implementing one or more of these functions and capabilities. Today's state-of-the-art computer technology provides means not available in the 1970s to implement intelligent retrieval systems. New distributed system architectures, processing equipment and configurations (including intelligent, microcomputer-based user workstations), and software languages and techniques are supporting the implementation of natural language query input and linguistic analysis of that input, graphic aids to browsing, closest-match, probabilistic retrieval methods (weighted term/logic and ranked output), and sophisticated user interface dialog/display techniques to engage the searcher intelligently for purposes of acquiring relevance feedback and search strategy modification.

## 5. Intelligent Interfaces and Intelligent Retrieval Systems or Software Techniques: A Systematic Review

For reasons of economy and complexity management, the illustrated review of intelligent information retrieval (IIR) systems and software which follows is both selective and classified. The conceptual ground at the cutting edge of developments is unsettled and uneven, and the existing accounts of major issues being addressed has not yet produced a common, accepted universe of discourse. The *solutions* represented by new intelligent systems and software vary in scope and kind. A few IIR system comprehensive design and development efforts have produced complete, stand-alone, intelligent retrieval systems (no question of a mere "front end" here). Most developments have been aimed at one or two known problem areas associated with conventional retrieval systems. The classified approach adopted here will bring together for purposes of exposition various efforts which have common aims, if not common design solutions. Likewise, efforts that address different issues are separated accordingly.

Figure 21 identifies the information retrieval (IR) process problem and task areas addressed by today's intelligent information retrieval (IIR) design activities, considered in their entirety. The categories identified in Figure 21 define the four-part topical outline of this review: 1) efforts to improve system entry and orientation-control for the user (Use Problem Areas 1 & 2), 2) vocabulary control and liberation solutions (Use Problem Area 3), 3) the provision of improved browsing and navigation facilities (System/Use Problem Areas 3b, 5d, 6), and 4) the use of new "smarter" query formulation, search strategy selection, and retrieval techniques (System/Use Problem Areas 4c, 5d). It must be emphasized that in any discussion or analysis of efforts to improve *interactive*, adaptive, cooperative IIR systems, design and use problems or solutions are different sides of the same coin. Success on either side, system side or use/user side, requires an understanding of the problems and real possibilities for improvement which exist on the other side. This imperative stands in opposition to a prevalent system design perspective: namely, that the system "internals" can and should be constructed first; the "requirements" of the user interface can be handled later, almost as an afterthought. This author has seen too many bad query-language, indexing, and retrieval design implementations that could not be improved or glossed-over with any amount of "front-end" software.

A major issue which cuts across all IIR research and development activities is the issue of control and delegation of roles and task accomplishment between the system and the user. What should the system do fully automatically (and perhaps transparently) for the user, and where and how should the user be engaged for his input or decision-making? Smith discusses this issue in terms of machine-aided intelligence ("augmentation") and machine intelligence ("delegation"): "In augmentation the computer assists the user in the substance of his task; in delegation decisions are made by the system itself using programmed criteria."[51] The terms "assistance" and "consultation" are often used as synonymous with "augmentation"; "supplantation" may be used as a substitute for "delegation".

The issue of user/system control and delegation is not well understood, and it may ultimately be a philosophical issue. Fortunately, the issue is beyond

| Design Tasks: | User-System Request/ Presentation Interaction | WIND | Thesaurus/Index Lookup/Matching: content, structure, linkages | Auto. Retrieval Logic/Inference Operations: Bool. prob'listic, etc. | Search Results Output Prep. & Presentation: format, rank | Database Design, Creation, Maintenance |
|---|---|---|---|---|---|---|
| Use Tasks: | a. | OWS | b. | c. | d. | e. |
| 1. Entry Mechanics (getting started) | | | | | | |
| 2. Orientation/Help Assistance | | | | | | |
| 3. Vocabulary Selection, Trans., Negotiation | | | ——— | - - - | ——— | |
| 4. Query Formulation, Modificat'n, Expans'n | | | ——— | ——— | ——— | |
| 5. Search Results Relevancy Assessment | | | - - - | - - - | ——— | |
| 6. Browse/Navigation | | | ——— | - - - | ——— | |

**Figure 21.** *The Information Retrieval Process: Problem/Task Areas*

the scope of this study. The trend in expert systems and intelligent retrieval is to combine approache; in a shared decision-making environment which includes automatic system tasking and meaningful engagement of the user at key points in the retrieval process cycle. (See Figures 22 and 23) However, the fact that some design developments and methods in IIR are transparent (i.e., invisible) to the user explains the impossibility of directly illustrating them in this report.

Commercially-available "front ends" for database access and searching are not reviewed here because, with one very recent exception (the TOME Searcher), these products are not very intelligent. Vendors may claim their products, generally available for microcomputers, assist the user in the various steps of the online search process by incorporating the expertise of trained, experienced search intermediaries. In fact, most of them are limited to providing assistance in areas not critical in OPAC subject searching (e.g., automatic dial-up and logon, database selection, and query translation into the host system's command language).[52] Most front ends go no further than automating the routing and query language conversion functions carried out by intermediaries.[53] As Doszkocs explains: "Generally, only minimal automatic assistance capabilities exist for search strategy formulation and vocabulary control ... Attempts to incorporate the expertise of trained searchers into the IR user interface are still in their infancy."[54] In addition, this approach, which puts the prevailing intermediary model "in" the system, may not be the best or most appropriate solution to the problems and needs of end-user online subject searching. Given the variety of search styles, aims, and requirements which exist, other models of search activity may be equally appropriate for inclusion and support in automated, interactive IIR systems.

## 5.1 Information Retrieval/OPAC System Entry and Orientation Assistance

Using the naval metaphor of "docking", Bates explains that "Any time a person approaches an information retrieval system,...an initial phase of orientation or 'getting a feel for' must be gone through before settling down to do searching proper."[55] These entry (getting started) and orientation needs are of two types: 1) "How do I use this system to start my search?" (= the mechanics of system use and searching), and 2) "What's going on now?" "I forget what I've done." "What can I do now?" "Why did *that* happen?" (= conceptual and physical orientation: knowing where one is in relationship to the system environment, all of which is not visible; not feeling confused or disoriented). These system use needs may translate into the unexamined popular notions "friendly" and "ease of use."

Advanced OPACs in operation are doing perhaps their best job in this problem/need area. Designers have employed various techniques for simplifying the learning and use of system search mechanics, and for keeping the searcher informed, oriented, and "in context." Search selection, search entry, and search expansion are eased in present-day OPACs through menu selection, reduced

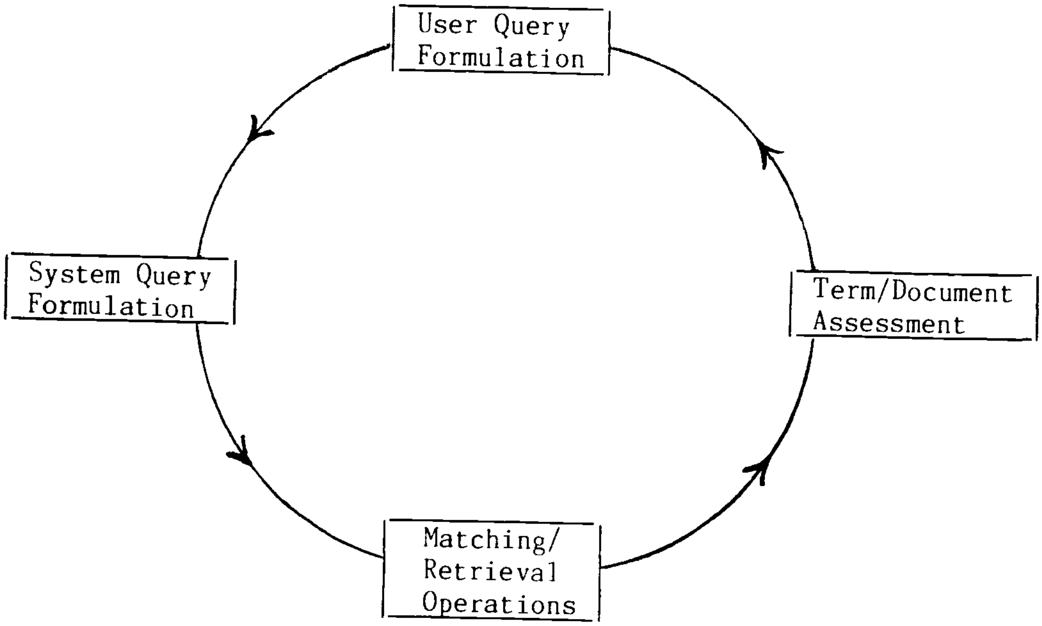


**Figure 22.** *Automated Information Retrieval Cycle*

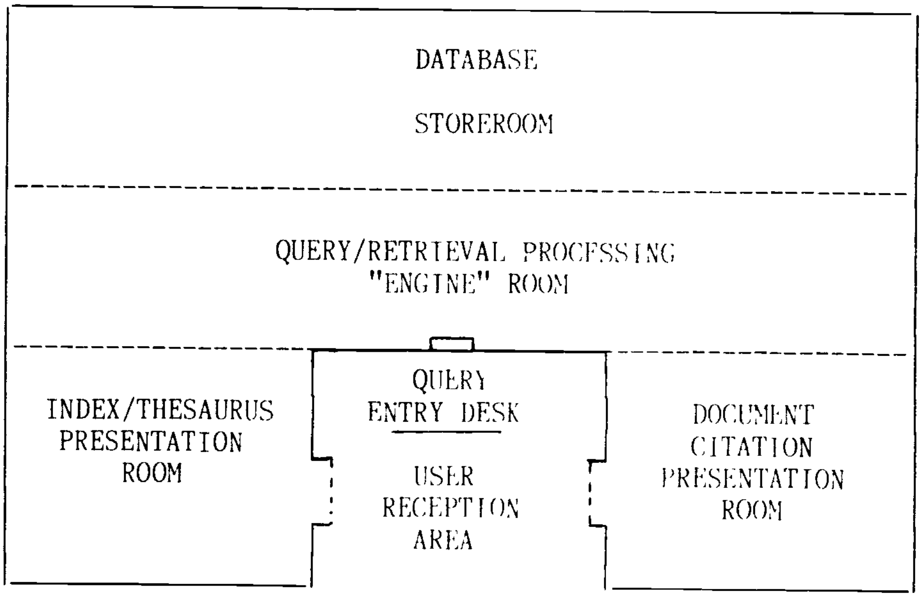


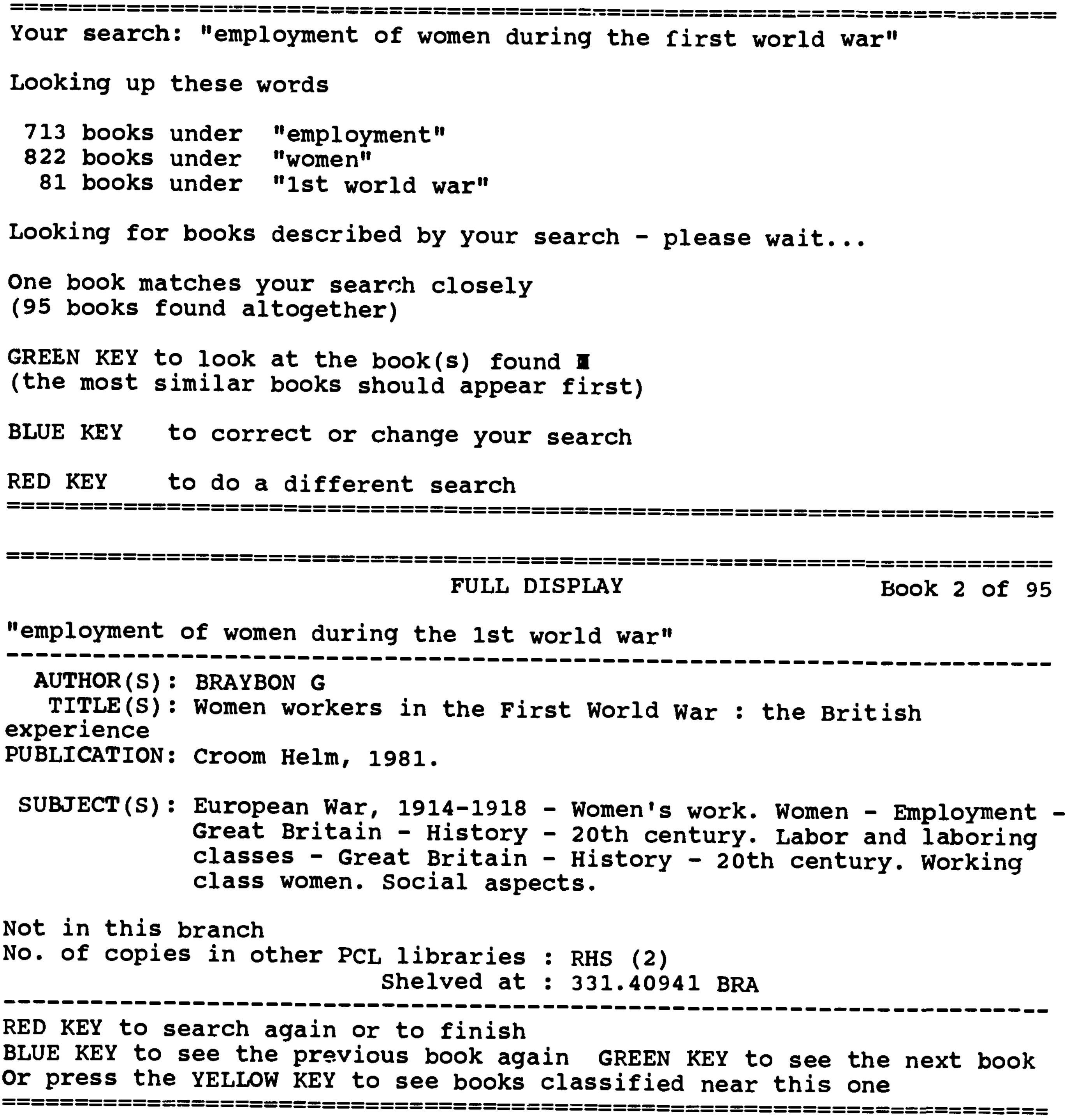
**Figure 23.** *Document "Retrieval House"*

command syntax requirements, or conversational question-answer user-system dialogue. The physical mechanics required for user selection and response actions are facilitated through the use of marked or coded function keys and keys to control on-screen cursor movements. Many of the complex query syntax and entry requirements present in conventional information retrieval systems have been reduced to "point" and "press" actions. Query formulation and articulation have been eased through the system's assumption of implicit search and display commands, automatic "insertion" of Boolean, truncation or proximity operators, and graphical query input aids.

The Dartmouth College and PALS OPAC search interfaces have already been illustrated, but, to review, the Dartmouth OPAC - which includes a software front end to BRS/Search - provides menus for function/search type selection and automatically constructs the Boolean query to be performed, requiring the user only to enter search terms. Search field(s) selection is simplified by their consolidation into menu options such as "Topic", "General" , "Author", and "Title" (See Figure 11). PALS accepts a simple two-character command for the selection of search type, then, as in the Dartmouth OPAC, automatically constructs the formal query. PALS goes a little further by prompting for search delimiters when too many records are retrieved, and by suggesting the use of the general keyword search strategy when the search just performed using another strategy results in no matches (See Figure 8).

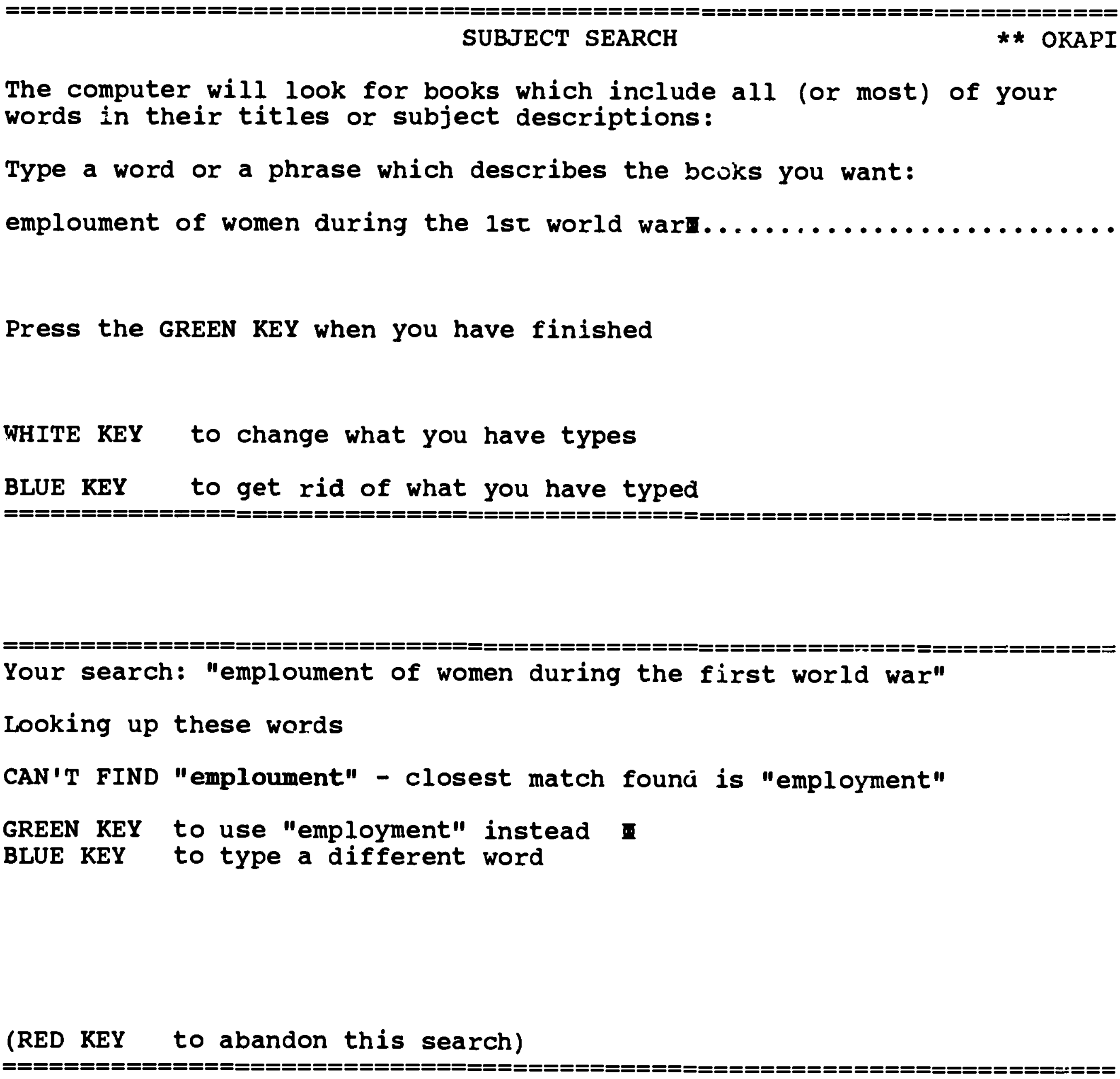
Figure 24 illustrates the non-traditional search entry and search interaction dialogue techniques used in OKAPI, the experimental/demonstration OPAC developed by a research team at the Polytechnic of Central London. OKAPI is placed in one of the Polytechnic's library sites for ongoing evaluation. After selecting the subject search option from the initial menu, the user is requested to enter his search in free-form natural language. OKAPI performs a non-Boolean, "closest match" search on user queries, but, first, in the example shown, a possible misspelling of a search word is displayed ("emploument"). OKAPI offers a correction in the list of terms for which it then searches. OKAPI also utilizes color-coded function keys which are referenced on screen to direct special user-invoked requests or response actions.

BiblioFile's "Intelligent Catalog" search/display interface provides menu, cursor-pointer selection techniques, user-requested help message windows, natural language search statement input, and automatic query formulation (See Figure 25a-f). When the user enters a compound statement and the phrase does not appear in the indexes, the system automatically constructs and performs a Boolean AND query. Screen-partitioning techniques bring a dynamic list of matching index entries to a portion of the screen; the entries displayed may change as the user modifies, shortens, or extends his search term(s). Search status, postings data, and search term field occurrence information are displayed to the searcher at appropriate stages of the search.



```
===================================================================
Your search: "employment of women during the first world war"

Looking up these words

 713 books under  "employment"
 822 books under  "women"
  81 books under  "1st world war"

Looking for books described by your search - please wait...

One book matches your search closely
(95 books found altogether)

GREEN KEY to look at the book(s) found ▮
(the most similar books should appear first)

BLUE KEY   to correct or change your search

RED KEY    to do a different search
===================================================================

===================================================================
                              FULL DISPLAY            Book 2 of 95

"employment of women during the 1st world war"
-------------------------------------------------------------------
  AUTHOR(S): BRAYBON G
   TITLE(S): Women workers in the First World War : the British
experience
PUBLICATION: Croom Helm, 1981.

 SUBJECT(S): European War, 1914-1918 - Women's work. Women - Employment -
             Great Britain - History - 20th century. Labor and laboring
             classes - Great Britain - History - 20th century. Working
             class women. Social aspects.

Not in this branch
No. of copies in other PCL libraries : RHS (2)
                          Shelved at : 331.40941 BRA
-------------------------------------------------------------------
RED KEY to search again or to finish
BLUE KEY to see the previous book again  GREEN KEY to see the next book
Or press the YELLOW KEY to see books classified near this one
===================================================================
```

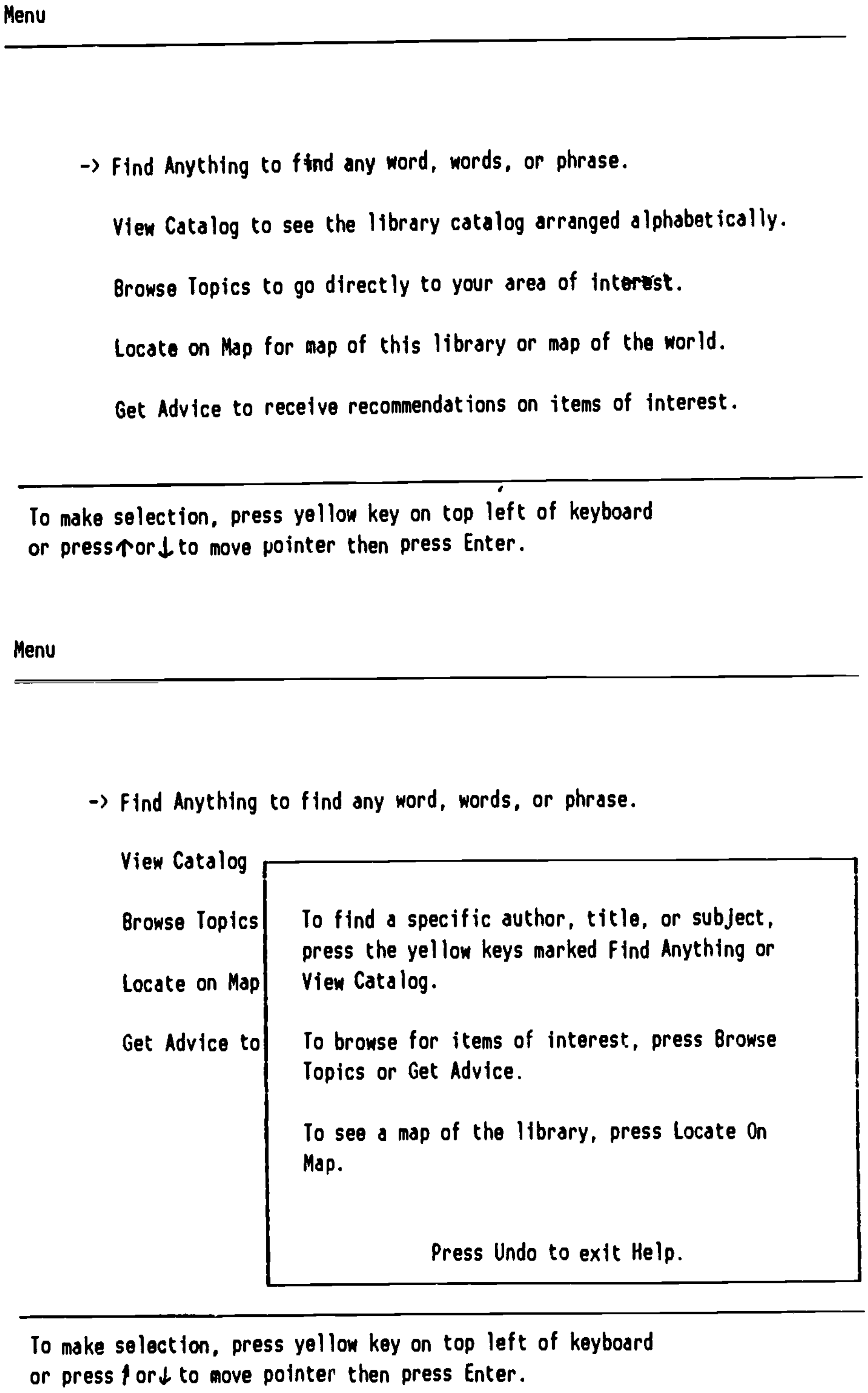
**Figure 24.** *OKAPI, Polytechnic University of Central London*



```
========================================================================
                              SUBJECT SEARCH                    ** OKAPI

The computer will look for books which include all (or most) of your
words in their titles or subject descriptions:

Type a word or a phrase which describes the books you want:

emploument of women during the 1st world war■...........................


Press the GREEN KEY when you have finished


WHITE KEY   to change what you have types

BLUE KEY    to get rid of what you have typed
========================================================================
```

```
========================================================================
Your search: "emploument of women during the first world war"

Looking up these words

CAN'T FIND "emploument" - closest match found is "employment"

GREEN KEY  to use "employment" instead  ■
BLUE KEY   to type a different word



(RED KEY   to abandon this search)
========================================================================
```

**Figure 24.** *OKAPI (Continued)*



Menu

-> Find Anything to find any word, words, or phrase.

View Catalog to see the library catalog arranged alphabetically.

Browse Topics to go directly to your area of interest.

Locate on Map for map of this library or map of the world.

Get Advice to receive recommendations on items of interest.

To make selection, press yellow key on top left of keyboard
or press↑or↓to move pointer then press Enter.

Menu

-> Find Anything to find any word, words, or phrase.

View Catalog

Browse Topics

Locate on Map

Get Advice to

To find a specific author, title, or subject, press the yellow keys marked Find Anything or View Catalog.

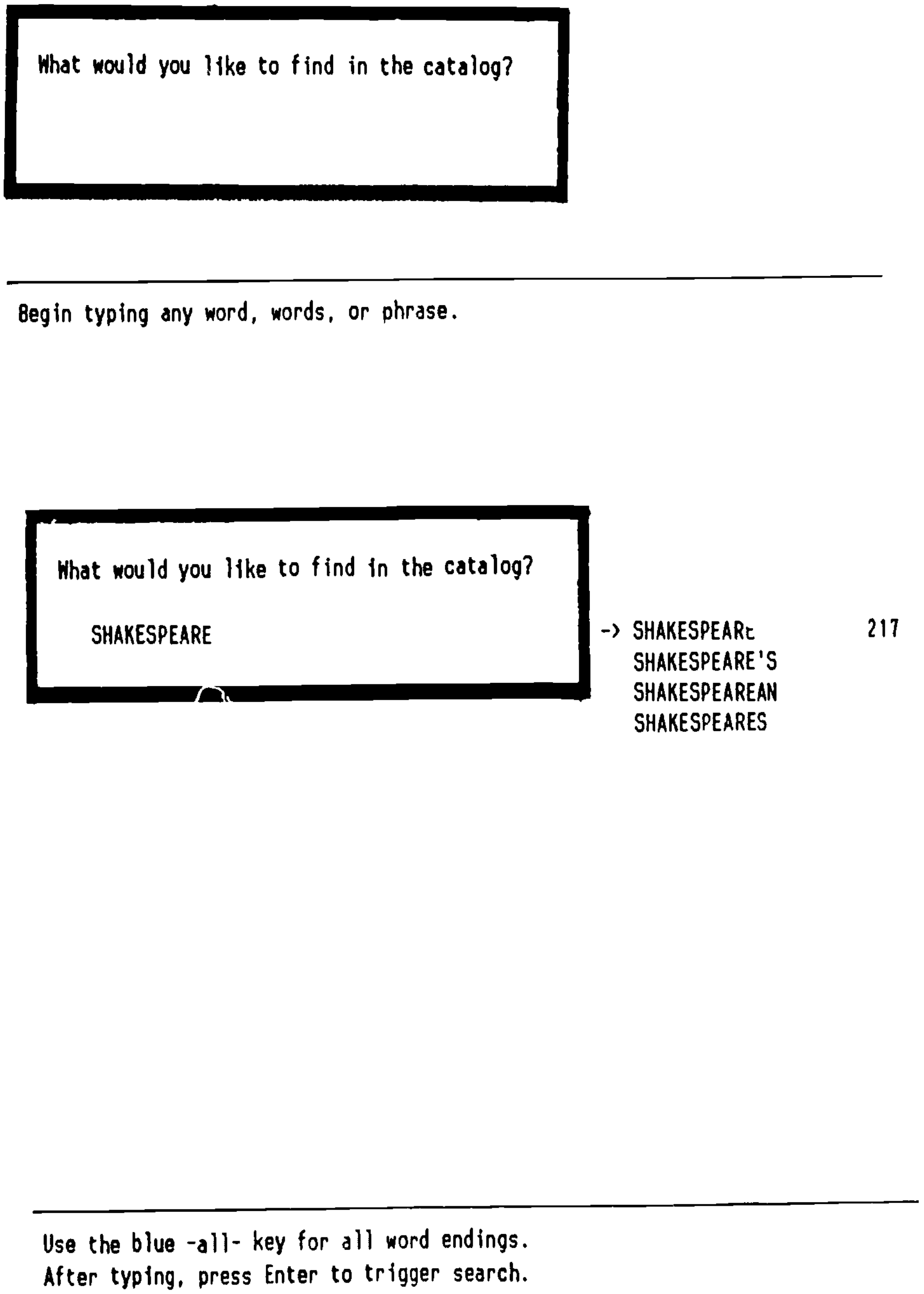
To browse for items of interest, press Browse Topics or Get Advice.

To see a map of the library, press Locate On Map.

Press Undo to exit Help.

To make selection, press yellow key on top left of keyboard
or press ↑ or ↓ to move pointer then press Enter.

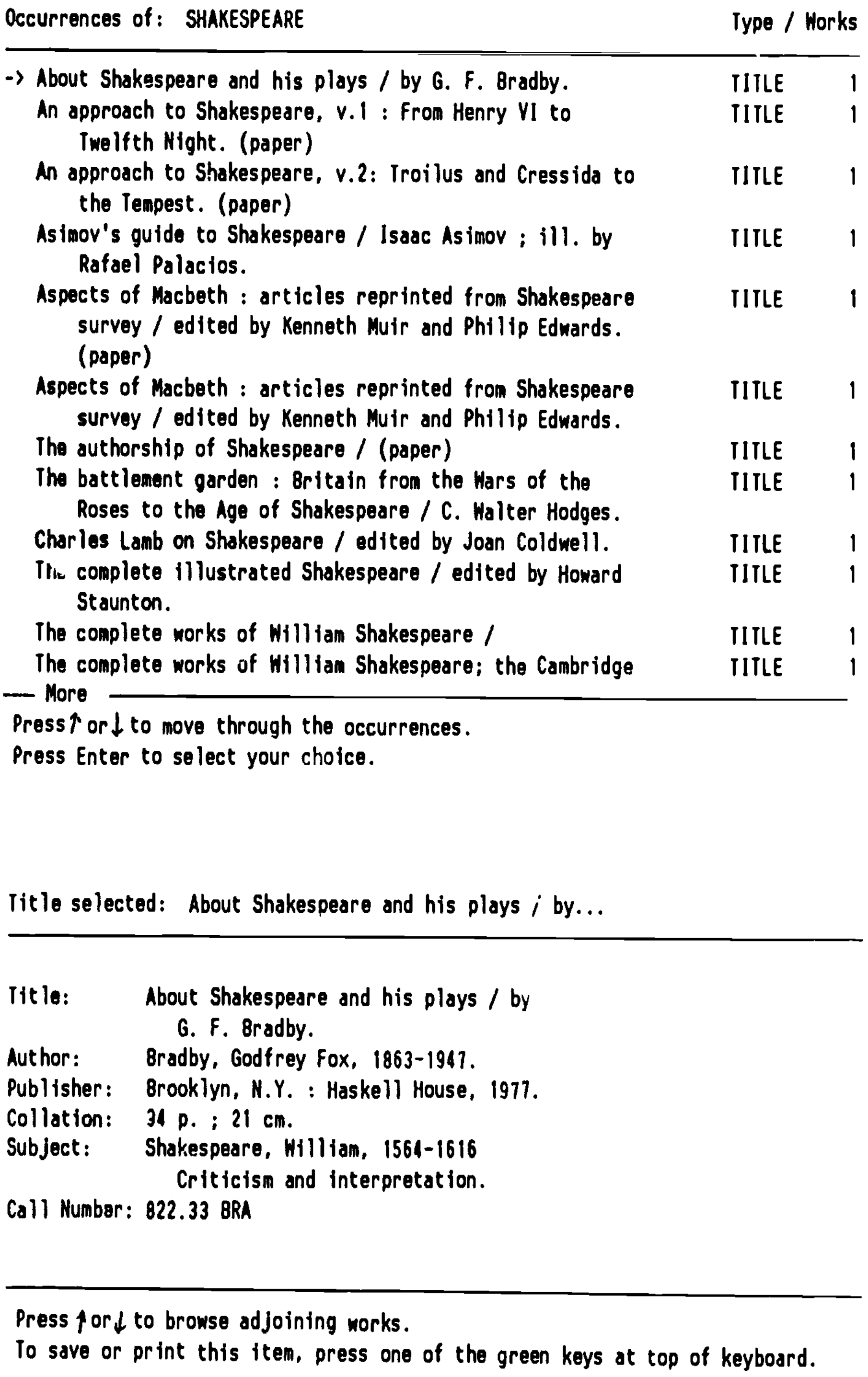
**Figure 25 a, b.** *BiblioFile Intelligent Catalog*



```
What would you like to find in the catalog?

Begin typing any word, words, or phrase.
```

```
What would you like to find in the catalog?

    SHAKESPEARE                                  -> SHAKESPEARE          217
                                                    SHAKESPEARE'S
                                                    SHAKESPEAREAN
                                                    SHAKESPEARES

Use the blue -all- key for all word endings.
After typing, press Enter to trigger search.
```

**Figure 25 c, d.** *Bibliofile Intelligent Catalog*



```
Occurrences of:  SHAKESPEARE                                        Type / Works
________________________________________________________________________________
-> About Shakespeare and his plays / by G. F. Bradby.              TITLE     1
   An approach to Shakespeare, v.1 : From Henry VI to              TITLE     1
      Twelfth Night. (paper)
   An approach to Shakespeare, v.2: Troilus and Cressida to        TITLE     1
      the Tempest. (paper)
   Asimov's guide to Shakespeare / Isaac Asimov ; ill. by          TITLE     1
      Rafael Palacios.
   Aspects of Macbeth : articles reprinted from Shakespeare        TITLE     1
      survey / edited by Kenneth Muir and Philip Edwards.
      (paper)
   Aspects of Macbeth : articles reprinted from Shakespeare        TITLE     1
      survey / edited by Kenneth Muir and Philip Edwards.
   The authorship of Shakespeare / (paper)                         TITLE     1
   The battlement garden : Britain from the Wars of the            TITLE     1
      Roses to the Age of Shakespeare / C. Walter Hodges.
   Charles Lamb on Shakespeare / edited by Joan Coldwell.          TITLE     1
   Th. complete illustrated Shakespeare / edited by Howard         TITLE     1
      Staunton.
   The complete works of William Shakespeare /                     TITLE     1
   The complete works of William Shakespeare; the Cambridge        TITLE     1
-- More ________________________________________________________________________
 Press ↑ or ↓ to move through the occurrences.
 Press Enter to select your choice.
```

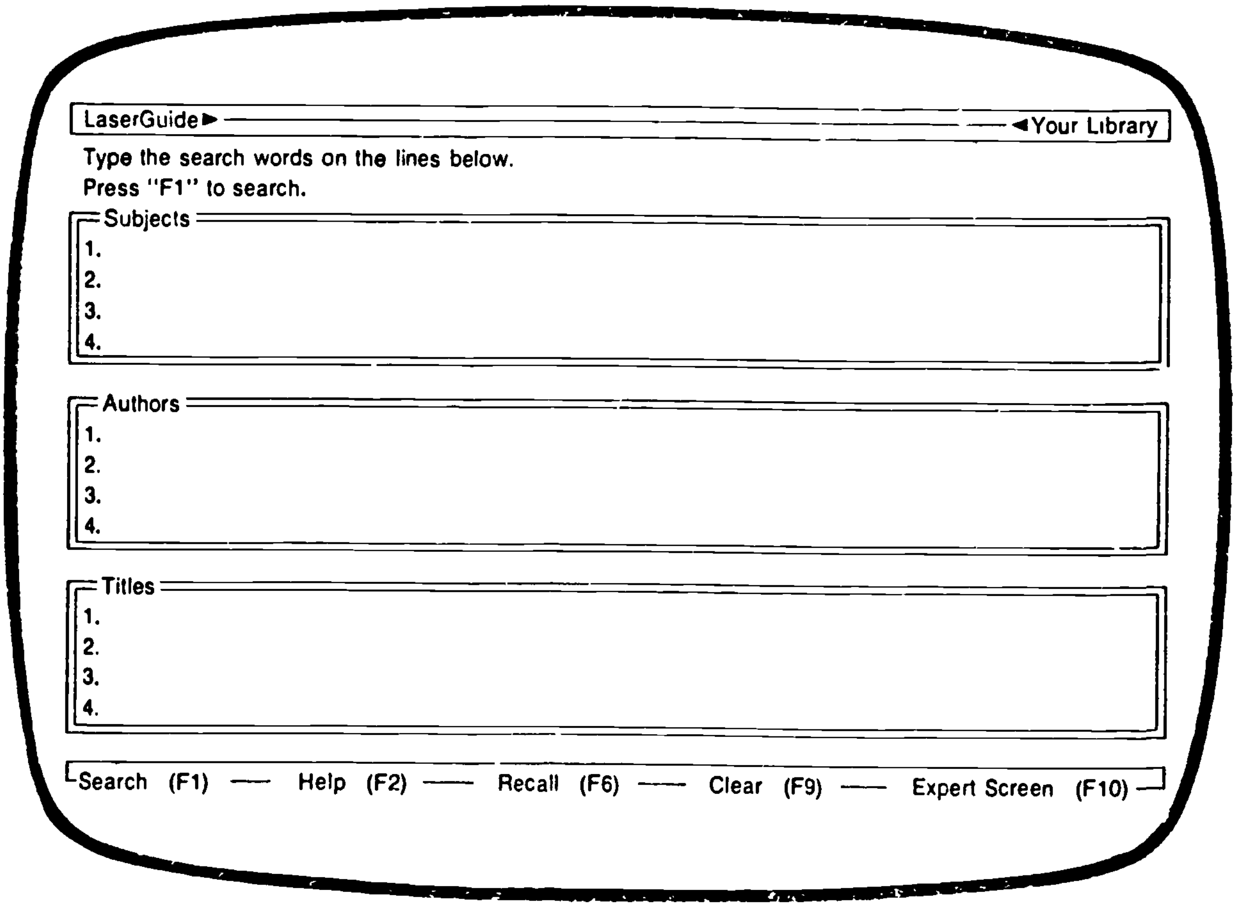
```
Title selected:  About Shakespeare and his plays / by...
________________________________________________________________________________

Title:        About Shakespeare and his plays / by
                 G. F. Bradby.
Author:       Bradby, Godfrey Fox, 1863-1947.
Publisher:    Brooklyn, N.Y. : Haskell House, 1977.
Collation:    34 p. ; 21 cm.
Subject:      Shakespeare, William, 1564-1616
                 Criticism and interpretation.
Call Number: 822.33 BRA

________________________________________________________________________________
 Press ↑ or ↓ to browse adjoining works.
 To save or print this item, press one of the green keys at top of keyboard.
```

**Figure 25 e, f.** *BiblioFile Intelligent Catalog*

TINlib's easy entry "Query-by-Form" and "Query-by-Example" screen templates have already been illustrated (See Figure 13). A similar approach is used by GRC's LaserGuide OPAC to simplify query input by the user (Figure 26a). Window boxes are used to direct search term entry; one box is designated for each type of search. Prompts at the bottom of each screen define special purpose function keys. These prompts are akin to road signs: we ignore them when we do not need them. LaserGuide's "Expert Search Screen" (Figure 26b) represents a novel design attempt to make Boolean query formulation easier and more intuitively understandable for the end user/searcher. Annotated "Boolean boxes" (the author's term) invite easy search term input. Another window box on the same screen may be used to enter search limiting criteria such as a date range, and function keys which permit search field selection are defined onscreen.

The Colorado Alliance of Research Libraries' (CARL) OPAC, marketed by The Eyring Research Institute, provides an informal, conversational style search interface (See Figures 27a-d). It presents a search as a seamless, continuing interactive process rather than as a disjointed series of discrete screens. In the fullest version of the CARL OPAC, the search options offered are Word (W), Name (N), and Browse (B). "B" leads to a choice of browsing lists (e.g., title, call number, etc.). Name and word search queries are processed automatically as keyword, Boolean AND searches. When CARL is used to support several types of databases or serves as a union library catalog, a simple "S" command invokes the automatic reinitiation of a query in another database or catalog selected by the user. When postings thresholds are reached in the matching process, the system invites the user to make the search more specific "by adding another word to your search." (Figure 27c) The CARL OPAC displays the number of matches for each search word and the combined number of hits for the Boolean AND ("+") operation. Browsing and "quick search" approaches may be used as alternatives to the basic system-guided, semi-automatic keyword-Boolean mode.

Another CD-ROM OPAC/IR system demands our attention (See Figure 28) because it employs interface techniques familiar to millions of business and consumer users of popular microcomputer software products. Bowker's CD-ROM retrieval system provides conventional Boolean retrieval as well as several index browsing search options. This system interface is notable for its special purpose, overlapping windows, window boxes, and pull-down menus used to present to the searcher multiple action options, search results screens, and status messages in a variety of configurations designed to maintain a sense of location and context in the overall search process. This approach helps keep the user oriented within the user-system environment. In "Browse" mode, the searcher may select type of entry from a pull-down menu; this action calls up a index browsing window which replaces the "Search Workspace" window box displayed in "Search" mode. To make specific selections or invoke special actions the searcher uses cursor keys to move a light bar and/or presses a defined function key. In the Bowker system and most other commercially-supplied systems, many of the search and display features can be customized to satisfy local needs and preferences.



LaserGuide ► ◄ Your Library

Type the search words on the lines below.
Press "F1" to search.

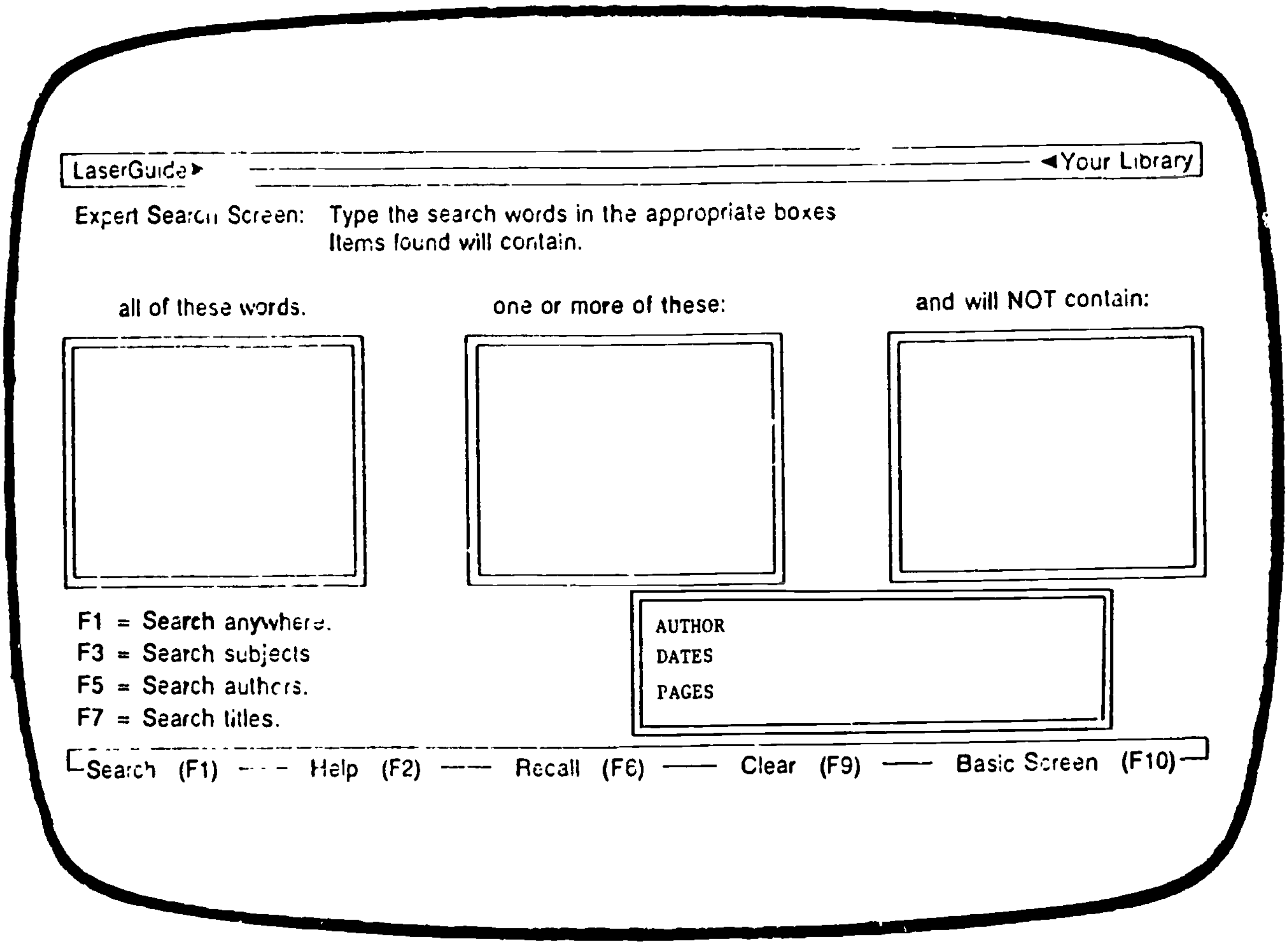
Subjects
1.
2.
3.
4.

Authors
1.
2.
3.
4.

Titles
1.
2.
3.
4.

Search (F1) — Help (F2) — Recall (F6) — Clear (F9) — Expert Screen (F10)

**Figure 26 a.** *LaserGuide Basic Search Screen*



LaserGuide ►
◄Your Library

Expert Search Screen: Type the search words in the appropriate boxes
Items found will contain.

all of these words.

one or more of these:

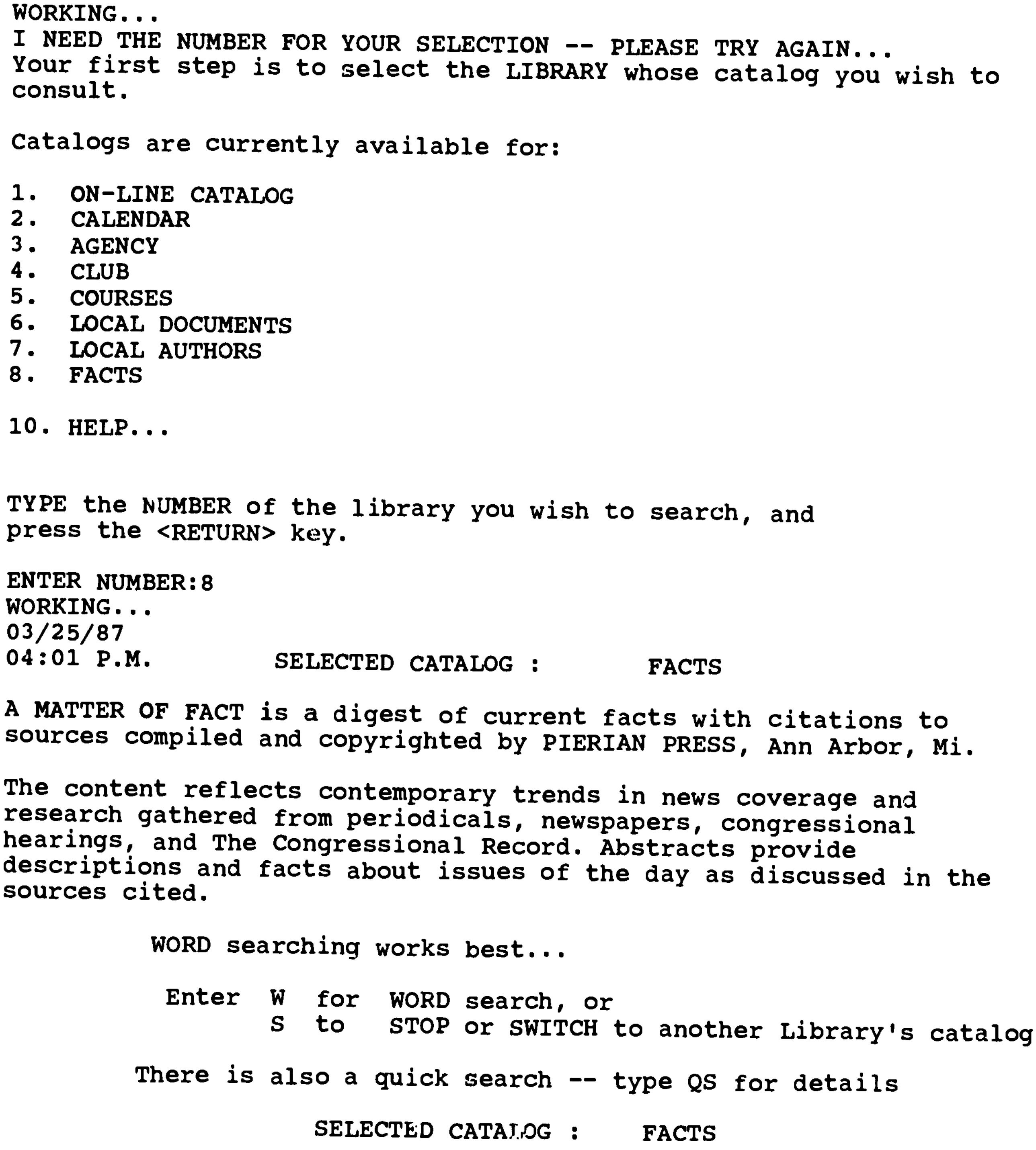
and will NOT contain:

F1 = Search anywhere.
F3 = Search subjects
F5 = Search authors.
F7 = Search titles.

AUTHOR
DATES
PAGES

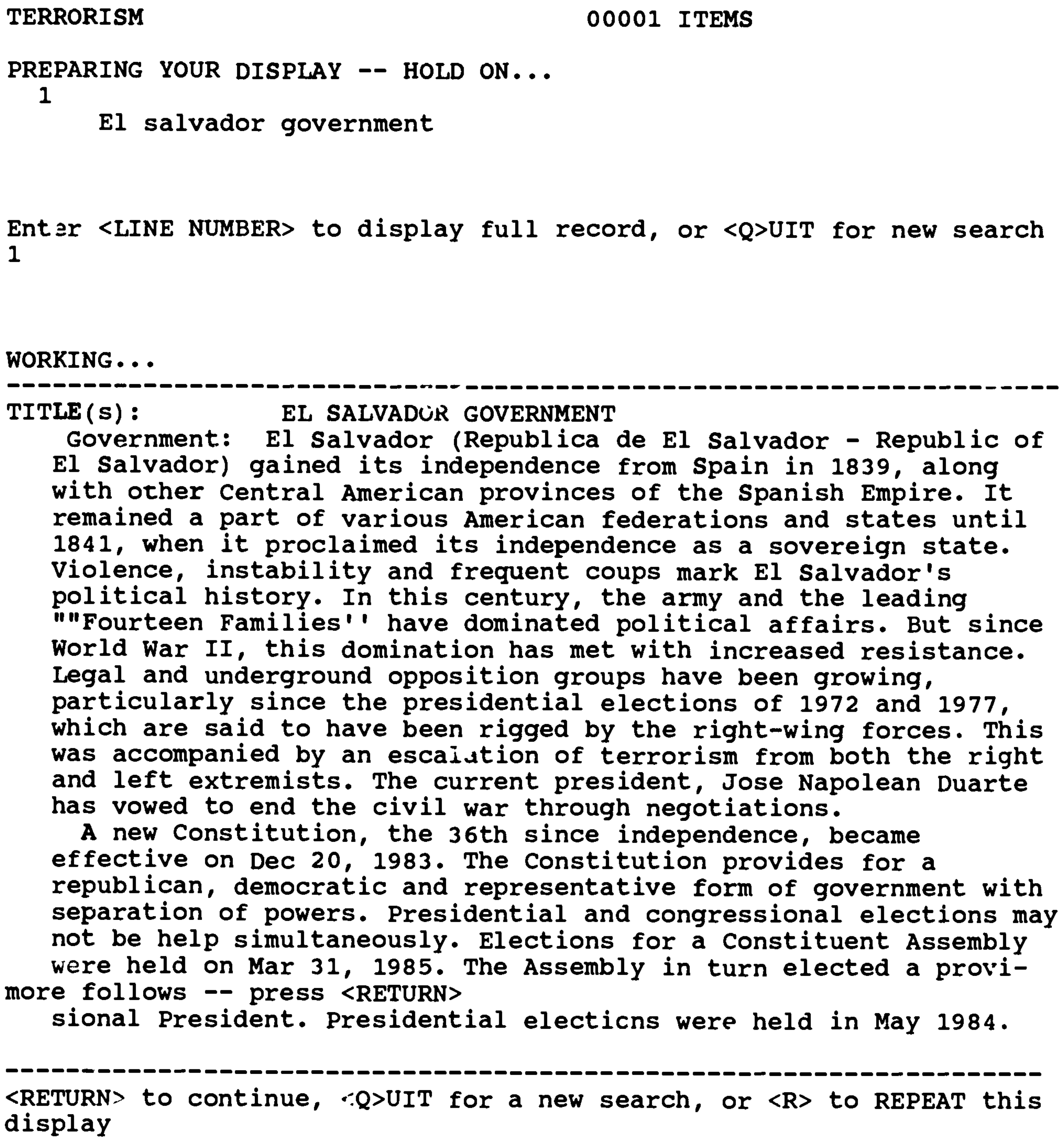
Search (F1) --- Help (F2) --- Recall (F6) --- Clear (F9) --- Basic Screen (F10)

**Figure 26 b.** *LaserGuide Expert Search Screen*



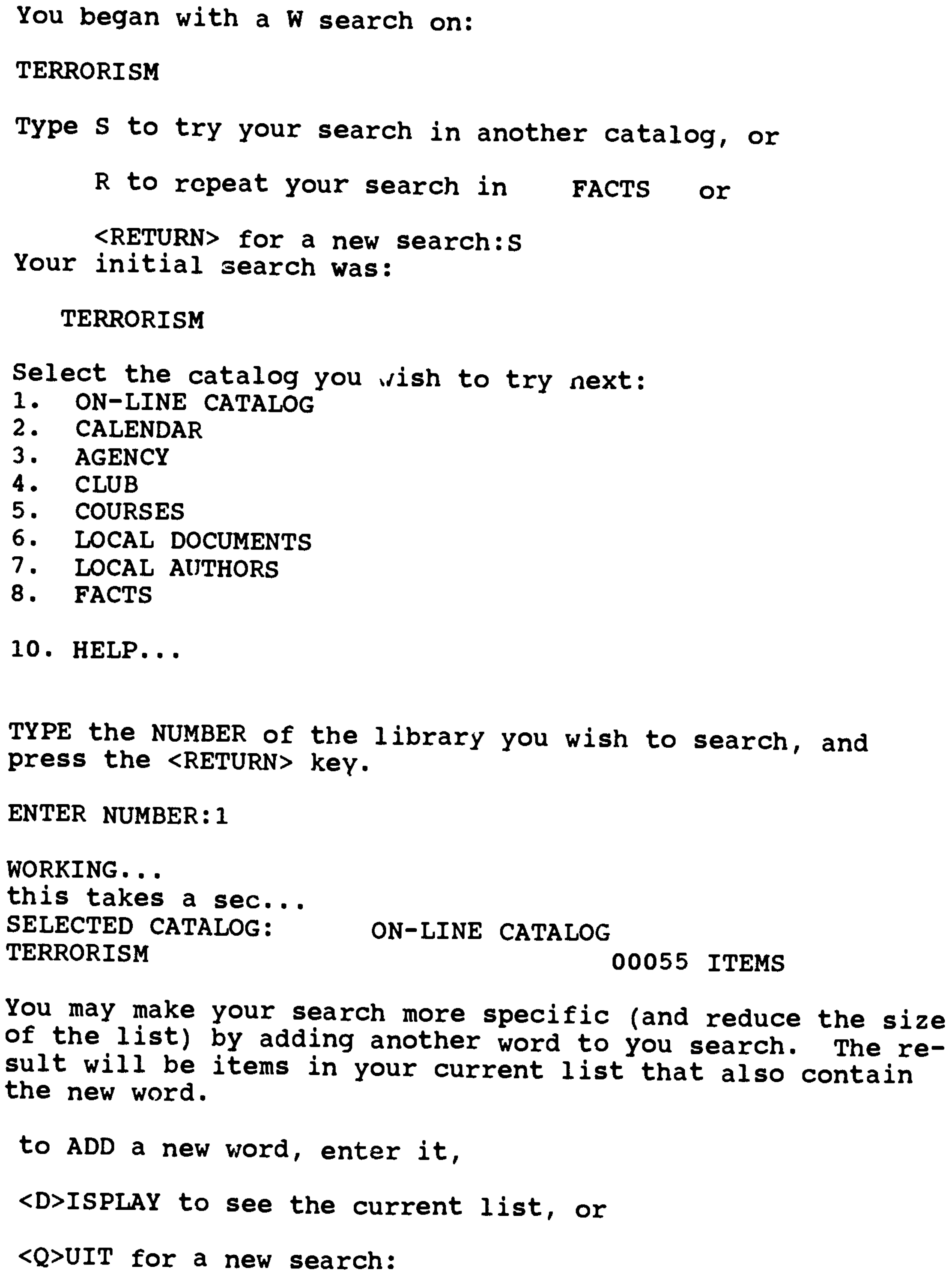
```
WORKING...
I NEED THE NUMBER FOR YOUR SELECTION -- PLEASE TRY AGAIN...
Your first step is to select the LIBRARY whose catalog you wish to
consult.

Catalogs are currently available for:

1.  ON-LINE CATALOG
2.  CALENDAR
3.  AGENCY
4.  CLUB
5.  COURSES
6.  LOCAL DOCUMENTS
7.  LOCAL AUTHORS
8.  FACTS

10. HELP...


TYPE the NUMBER of the library you wish to search, and
press the <RETURN> key.

ENTER NUMBER:8
WORKING...
03/25/87
04:01 P.M.         SELECTED CATALOG :       FACTS

A MATTER OF FACT is a digest of current facts with citations to
sources compiled and copyrighted by PIERIAN PRESS, Ann Arbor, Mi.

The content reflects contemporary trends in news coverage and
research gathered from periodicals, newspapers, congressional
hearings, and The Congressional Record. Abstracts provide
descriptions and facts about issues of the day as discussed in the
sources cited.

          WORD searching works best...

           Enter  W  for  WORD search, or
                  S  to   STOP or SWITCH to another Library's catalog

         There is also a quick search -- type QS for details

                     SELECTED CATALOG :    FACTS
```

**Figure 27 a.** *CARL: Pikes Peak Library*



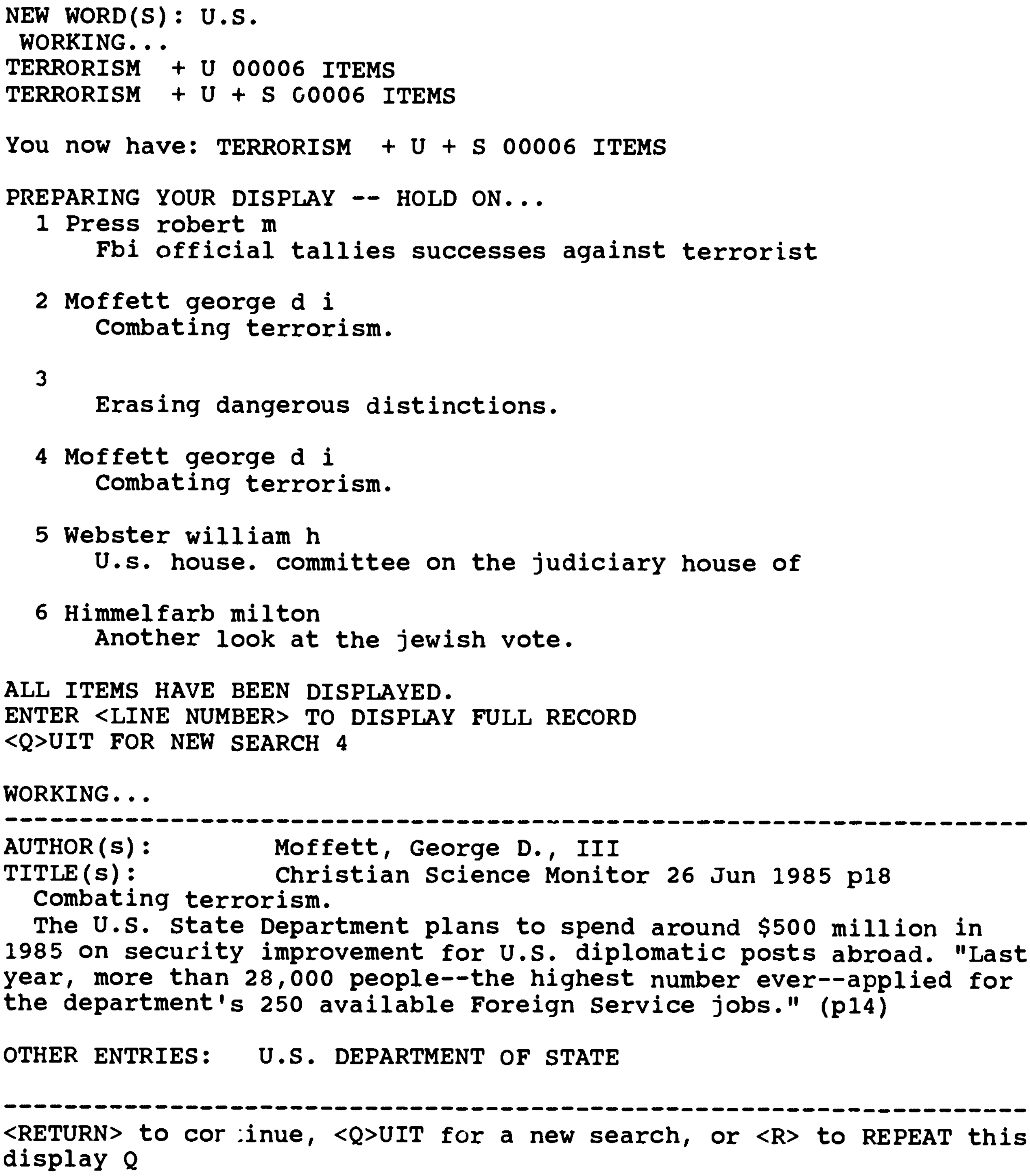
```
TERRORISM                                 00001 ITEMS

PREPARING YOUR DISPLAY -- HOLD ON...
  1
     El salvador government



Enter <LINE NUMBER> to display full record, or <Q>UIT for new search
1



WORKING...
---------------------------------------------------------------
TITLE(s):              EL SALVADOR GOVERNMENT
    Government:  El Salvador (Republica de El Salvador - Republic of
    El Salvador) gained its independence from Spain in 1839, along
    with other Central American provinces of the Spanish Empire. It
    remained a part of various American federations and states until
    1841, when it proclaimed its independence as a sovereign state.
    Violence, instability and frequent coups mark El Salvador's
    political history. In this century, the army and the leading
    ""Fourteen Families'' have dominated political affairs. But since
    World War II, this domination has met with increased resistance.
    Legal and underground opposition groups have been growing,
    particularly since the presidential elections of 1972 and 1977,
    which are said to have been rigged by the right-wing forces. This
    was accompanied by an escalation of terrorism from both the right
    and left extremists. The current president, Jose Napolean Duarte
    has vowed to end the civil war through negotiations.
      A new Constitution, the 36th since independence, became
    effective on Dec 20, 1983. The Constitution provides for a
    republican, democratic and representative form of government with
    separation of powers. Presidential and congressional elections may
    not be help simultaneously. Elections for a Constituent Assembly
    were held on Mar 31, 1985. The Assembly in turn elected a provi-
more follows -- press <RETURN>
    sional President. Presidential electicns were held in May 1984.

--------------------------------------------------------------
<RETURN> to continue, <Q>UIT for a new search, or <R> to REPEAT this
display
```

Figure 27 b. *CARL (Continued)*



```
You began with a W search on:

TERRORISM

Type S to try your search in another catalog, or

     R to rcpeat your search in    FACTS   or

     <RETURN> for a new search:S
Your initial search was:

   TERRORISM

Select the catalog you wish to try next:
1.  ON-LINE CATALOG
2.  CALENDAR
3.  AGENCY
4.  CLUB
5.  COURSES
6.  LOCAL DOCUMENTS
7.  LOCAL AUTHORS
8.  FACTS

10. HELP...


TYPE the NUMBER of the library you wish to search, and
press the <RETURN> key.

ENTER NUMBER:1

WORKING...
this takes a sec...
SELECTED CATALOG:      ON-LINE CATALOG
TERRORISM                             00055 ITEMS

You may make your search more specific (and reduce the size
of the list) by adding another word to you search.  The re-
sult will be items in your current list that also contain
the new word.

 to ADD a new word, enter it,

 <D>ISPLAY to see the current list, or

 <Q>UIT for a new search:
```

**Figure 27 c.** *CARL (Continued)*



```
NEW WORD(S): U.S.
 WORKING...
TERRORISM  + U 00006 ITEMS
TERRORISM  + U + S C0006 ITEMS

You now have: TERRORISM  + U + S 00006 ITEMS

PREPARING YOUR DISPLAY -- HOLD ON...
  1 Press robert m
      Fbi official tallies successes against terrorist

  2 Moffett george d i
      Combating terrorism.

  3
      Erasing dangerous distinctions.

  4 Moffett george d i
      Combating terrorism.

  5 Webster william h
      U.s. house. committee on the judiciary house of

  6 Himmelfarb milton
      Another look at the jewish vote.

ALL ITEMS HAVE BEEN DISPLAYED.
ENTER <LINE NUMBER> TO DISPLAY FULL RECORD
<Q>UIT FOR NEW SEARCH 4

WORKING...
--------------------------------------------------------------------
AUTHOR(s):         Moffett, George D., III
TITLE(s):          Christian Science Monitor 26 Jun 1985 p18
  Combating terrorism.
  The U.S. State Department plans to spend around $500 million in
1985 on security improvement for U.S. diplomatic posts abroad. "Last
year, more than 28,000 people--the highest number ever--applied for
the department's 250 available Foreign Service jobs." (p14)

OTHER ENTRIES:   U.S. DEPARTMENT OF STATE

--------------------------------------------------------------------
<RETURN> to cor :inue, <Q>UIT for a new search, or <R> to REPEAT this
display Q
```

**Figure 27 d.** *CARL (Continued)*

**Figure 28.** *Bowker's CD-ROM*

## 5.2 Vocabulary Control, Matching and Translation

An abundance of evidence and experience suggests that the *major* subject access and retrieval problem in today's OPACs is the "vocabulary" problem. No other issue is as central to retrieval performance and user satisfaction. The problem has many dimensions, but it can be defined in simple terms as the query expression, document term-matching problem.

OPAC subject searchers have the dual task of coming up with terms that express their information needs or search concepts, and then finding the best terms in the system's language for these concepts or needs. Apparently, the user-system term "matching" process is more often than not one of translation and/or negotiation. Research indicates that in present-day OPACs the "negotiation" process is largely one-sided and fails to produce a match up to eighty percent of the time in initial attempts. Subject searchers typically enter whatever term or terms come to mind, with little thought given to how their needs might be expressed in the vocabulary of the system, controlled or uncontrolled. On the whole they seem unaware that a special purpose vocabulary (e.g., LCSH) is used to describe the subjects of books. As Markey discovered, "whatever pops into their minds" is entered in their subject searches as a hopeful "shot in the dark" (Bates). Failure analysis studies thus far show that the majority of OPAC subject searches end in failure of one kind or another: failure to match anything, broad-match failure that retrieves too many items, and the failure to retrieve precisely what is needed or desired on a given topic (i.e., a little or a lot of highly focused, relevant material).

Given that most of this subject searching is done in today's MARC/LCSH-based subject OPACs, it is easy to assign blame. Current OPAC system design places the burden of term selection and matching almost entirely on the user. LCSH is the primary subject access and matching vocabulary used in OPACs. Neither the user nor LCSH is well-equipped to tackle this matching/translation task. LCSH provides little "entry" vocabulary ("see" references) to translate or link users' search terms, which may consist of unused or variant-form words or phrases, to the LCSH terms used to index documents. Due to its shallow, inconsistent syndetic structure, LCSH is inadequate for guiding the user, or for automatically mapping search terms, to related terms that may have been used to index relevant materials. Markey's SULIRS OPAC study revealed that if users' search terms that failed to match LCSH terms exactly were translated to LCSH cross references or were truncated to reduce morphological variations of the same word, the match rate would have increased only from 18 percent to 29 percent.[56] At any rate, no OPAC yet provides online access to the LCSH list itself, so it could be used for entry vocabulary matching or as a browsing thesaurus. In today's OPACs subject access index entries are derived by crude phrase or keyword extraction of data from the subject fields of individual catalog records. Some OPACs are beginning to add references from LCSH to the catalog entries, but this is still exceptional and online access to the LCSH list has not been provided.

On the other hand, the deficiencies associated with user search vocabulary are often the deficiencies of language itself. Natural language is loaded with ambiguity, vagueness, and syntactical variations and complexities. Controlled vocabularies are designed to help with these problems, but unavoidable inconsistency and semantic loss occur in the application of controlled subject vocabularies in the indexing of documents. Words may have different meanings in different fields of study (e.g., "china", "bonding"). Is the searcher who enters these words interested in ceramic engineering or sociobiological behavior in the Peoples' Republic?

To aid "best" query term selection and matching, some degree of "vocabulary control" is required. Homonyms, synonyms, and morphological or orthographical word variants must be normalized and brought under control. Related-term linkages must be established to ameliorate the language matching problem in information retrieval.

These vocabulary control functions can be performed in three ways: by human thesaurus builders and indexers, by computer programs, and by retrieval system users at the time of searching. In fact, in commercial online database searching intermediaries are trained to select the "correct" index terms and possible variant forms or synonyms and to form them into a query based on their knowledge of the structural and semantic properties of particular databases and document collections. Without much system help with vocabulary control the burden falls on the user to consider all relevant search terms, including synonyms and variant word forms or spellings.

Present-day OPACs delegate this function to the end user at the time of search. This is unfortunate because, as Svenonius notes, most "users do not have the verbal imagination to conceive of all possible search terms;" and also, they may not "have time to key in all possible search terms."[57]

Users infrequently match exactly the LCSH terms in the catalog, and they are generally not very good at predicting how concepts and subjects will be represented in the *uncontrolled* vocabulary of catalog records. In a study of query term, title term match attempts, Tague found that "when the words of a user query were compared with the title words of relevant journal articles, only 31.8% of the title-query words pairs were exact matches and an additional 6.5% close variants which might be retrieved, for example, by a truncated word search."[58]

Clearly, OPAC subject searchers need more help with vocabulary control then they are presently receiving. Truncation of search terms and subject index displays are not help enough. Help may come from two directions: linguistic analysis and computation performed by the OPAC system software, and improved thesaural (i.e., semantic) aids to be put at the disposal of both the system and the system user. The second of these is the joint responsibility of thesaurus builders/indexers and system designers. Effective term selection, term matching or translation, and relevant document retrieval in library catalogs requires good vocabulary control and online assistance. When the indexers and the system assume this responsibility, vocabulary *liberation* may result for OPAC users.

Bates remarks, "any reasonable English language word or phrase should get the searcher started and linked to explanatory, guiding information to assist in the search."[59] This point is expanded by Congreve: "The library situation is unique in attempting complex database retrieval while assuming no previous expertise on the part of the user ... A fuzzy query from the user must be translated to give an effective search path across the data. The user may need to r odify these pathways and so will need sophisticated browsing facilities, and an expert dialogue to prompt for alternative search strategies."[60]

Congreve's statement frames the two aspects of the OPAC vocabulary problem, and points to the three kinds of solutions which are being worked on by designers of intelligent systems. One aspect of the problem is the entry vocabulary translation and matching task. Much of this can be accomplished with automatic and semi-automatic computer techniques not yet employed in most OPACs. Providing online thesaurus exploration and related-term guidance aids represents another aspect of the vocabulary problem. OPAC/IR research and development work on these problems is taking place in three areas: 1) exploiting "front-end" vocabularies such as dictionaries, semantic networks, and subject thesauri to enrich and increase the "lead-in" system entry vocabulary, 2) testing and using algorithmic linguistic analysis and processing techniques, including vocabulary mapping or switching, and 3) the design of improved vocabulary exploration and negotiation facilities such as expert dialogue assistance, relevance feedback/query expansion/navigation options, and graphical displays of thesauri using AI workstations with high-resolution, graphics display capabilities.

## Enriching the Entry Vocabulary

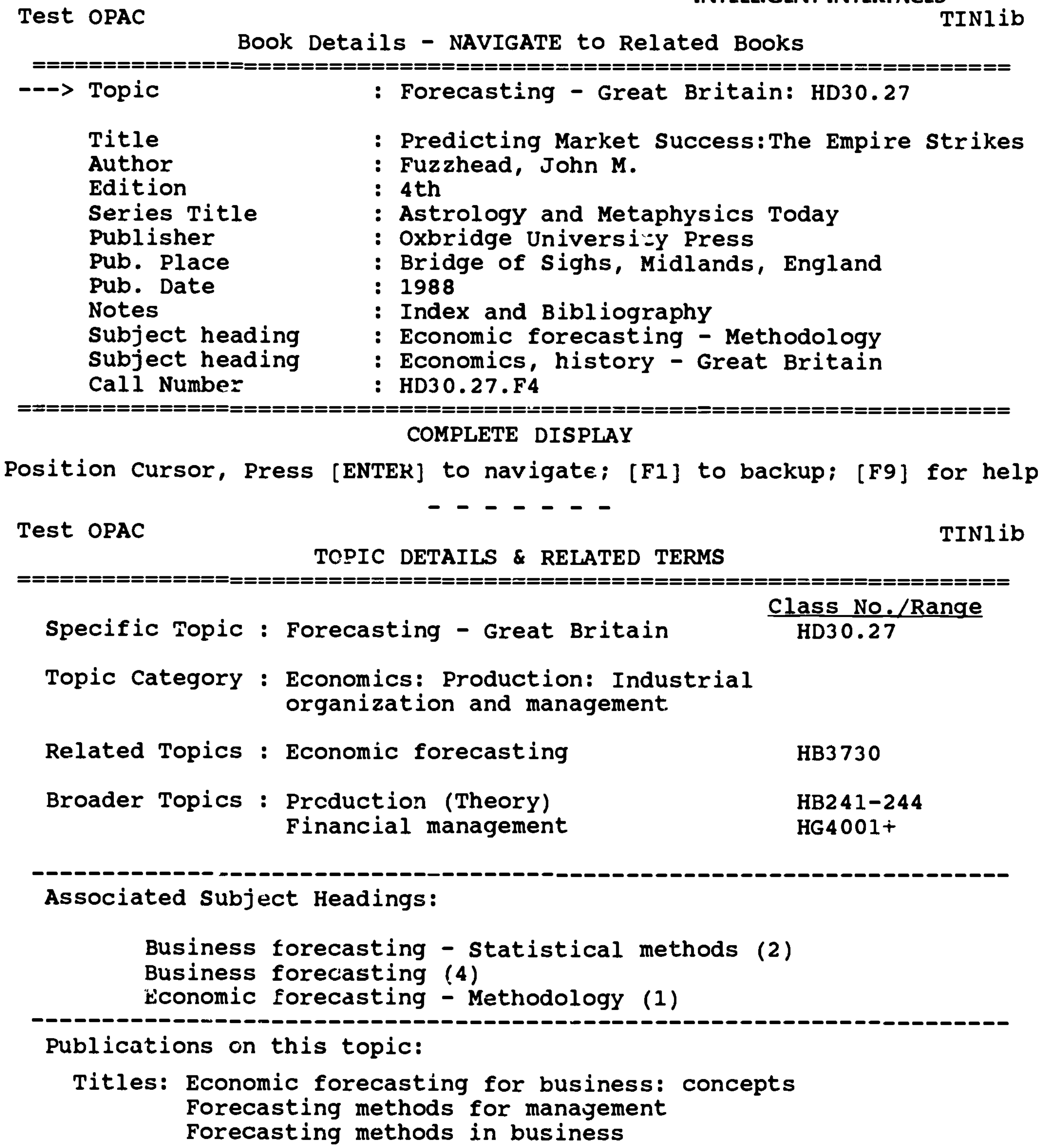
Many researchers have recommended adding terms from various external sources to OPAC subject indexes or to a front-end "lead-in" dictionary to those indexes, for the purpose of increasing initial search matching rates. Mandel and Herschman explain the motivation behind these suggestions: "Displaying the existing LCSH headings and references will not, in itself, solve the problem of an entry vocabulary that matches only half of users' first tries. Even without making a single change in an LC subject term, access to the terms could be improved enormously by adding to the entry vocabulary (i.e., adding "see" references)."[61] (*Note*: Studies indicate the matching rate is less than 30%.) In providing an enriched lead-in vocabulary to LCSH descriptors, or any other subject index scheme, two issues present themselves: 1) Where should these terms come from?, and 2) How are they to be linked to terms in the system's vocabulary? Library catalogs usually include records for materials in a large variety of subject fields and interest areas. Bringing all existing specialized subject thesauri to an OPAC front end is simply not fcasible. Of course, vocabulary sources internal to the database (e.g., title words or component LCSH words) have been exploited as "lead-in" vocabulary through keyword and rotated descriptor indexing techniques.

Usirg terms *implicit* in the catalog records, but registered in an external source (e.g., classification schedules), is a promising approach. It largely solves the formidable lead-in, system term linking problem, requiring little or no intellectual effort to create these "see" references. In most OPACs the thesauri used for subject cataloging are external to the OPAC database, as are the classification schedules and indexes. Terms from these sources can be added to the initial-search matching subject indexes in OPACs. Machine analysis can eliminate/merge most duplicates. Linkages between these new lead-in terms and subject headings or titles in catalog records can be established indirectly, but automatically, through their "co-occurrence" in individual records, even though that co-occurrence may be implicit and not normally visible to viewers of the records. Catalog records contain call numbers and other indicators of linkages to alternative subject vocabularies (e.g., MeSH, PRECIS, local headings, etc.).

Some librarians may worry about the lack of vocabulary control in machine-created lead-in vocabularies (variant word forms, homonyms, etc.). OPAC searchers would welcome this problem if it was accompanied by higher initial search term matching rates. It is not much of a problem in today's OPACs because there is so little lead-in, initial matching, vocabulary available. Control can be exercised more easily and cheaply by the system after the initial match occurs. Users can be shown related term contexts, or be taken directly to some matching records to begin the process of discrimination and focusing.

Markey's experiments with a classification (Dewey) vocabulary-enriched OPAC are well known.[62] Terms from the Dewey Decimal Classification (DDC) Schedules and Relative Index were added to the OPAC subject search indexes. When compared to a control OPAC which provided traditional subject searching capabilities, the Dewey-enhanced OPAC performed better in some areas. Initial query term matching was increased, and the Dewey OPAC (also enhanced with class schedule browsing facilities) searching produced both more successful searches than the control OPAC and searches that retrieved different sets of relevant records than the same searches in the control OPAC.

Hildreth is testing an experimental OPAC which exploits the vocabulary of the Library of Congress Classification (LCC) Schedules. Not previously used online, this vocabulary will be used to create an alternative subject entry index and a browsing thesaurus, to enrich the keyword index, and to provide explicit navigation linkages between LCSH headings and LCC descriptors (See Figure 29). Earlier studies have shown little overlap between LCC and LCSH terms, thus the entry vocabulary in this test OPAC should be enriched. It is suspected that LCC-based descriptors linked to LCSH headings will enable a searcher to focus his search on specific aspects or treatments of the problem, in effect,



```
Test OPAC                                                          TINlib
               Book Details - NAVIGATE to Related Books
=====================================================================
---> Topic                : Forecasting - Great Britain: HD30.27

     Title                : Predicting Market Success:The Empire Strikes
     Author               : Fuzzhead, John M.
     Edition              : 4th
     Series Title         : Astrology and Metaphysics Today
     Publisher            : Oxbridge University Press
     Pub. Place           : Bridge of Sighs, Midlands, England
     Pub. Date            : 1988
     Notes                : Index and Bibliography
     Subject heading      : Economic forecasting - Methodology
     Subject heading      : Economics, history - Great Britain
     Call Number          : HD30.27.F4
=====================================================================
                              COMPLETE DISPLAY
Position Cursor, Press [ENTER] to navigate; [F1] to backup; [F9] for help

                              - - - - - - -

Test OPAC                                                          TINlib
                    TOPIC DETAILS & RELATED TERMS
=====================================================================
                                                   Class No./Range
  Specific Topic : Forecasting - Great Britain       HD30.27

  Topic Category : Economics: Production: Industrial
                   organization and management

  Related Topics : Economic forecasting              HB3730

  Broader Topics : Production (Theory)               HB241-244
                   Financial management              HG4001+

  --------------------------------------------------------------------
  Associated Subject Headings:

         Business forecasting - Statistical methods (2)
         Business forecasting (4)
         Economic forecasting - Methodology (1)
  --------------------------------------------------------------------
  Publications on this topic:

    Titles: Economic forecasting for business: concepts
            Forecasting methods for management
            Forecasting methods in business
```

**Figure 29.** *Navigation Feature in TINlib*
*(From "Topic" to mixed thesauri, related terms display)*

*intelligently* narrowing his search. Current OPAC methods for narrowing or reducing search results are not semantic or subject-based (e.g., narrowing by date, language, or format). This experimental OPAC combines several different solutions to the vocabulary problem in subject searching: enriching entry vocabulary, linking different thesauri, and providing the "contextual" subject approach. The contextual subject approach involves displaying and guiding users through the system's vocabulary in its actual use for collocating and discriminating the variety of materials in a subject area of the collection.

In Great Britain, an OPAC research project at the Middlesex Polytechnic University is testing the use of the PRECIS thesaurus as a front end to a MARC database. The machine-readable PRECIS RIN file will be used for providing subject entry vocabulary, vocabulary control, and browsing displays. Assigned PRECIS headings are stored in the 690 fields of the UK MARC catalog records. However, the PRECIS RIN file will serve as a "buffer" between the searcher and the 690 subject descriptors. Congreve explains, "The object is to modify the user's input, if necessary, to match the controlled vocabulary of the PRECIS descriptors, and also, to provide some structured browsing facilities. If the user has generated too many items in the initial search, the system could suggest alternative, more specific, search terms. A search that has become too narrow, giving the user few or no references, needs to be broadened."[63]

Making a sound argument for system-generated vocabulary control and the provision of related terms, Congreve explains how this will be implemented in the experimental OPAC:

> The PRECIS RIN data consists, for each index term, of a unique identifying number (the Reference Indicator Number), a verbal statement of the term, and a list of the number of related terms, each being preceded by an operator which specifies the nature of this relationship.... Using these links it is possible to route an online catalogue user along the pathways of related terms, browsing down a hierarchically related sequence, for example, until the appropriate specific term is reached, as shown in [Figure 30].

The terms users input in successful subject searches is another source of entry vocabulary not yet exploited in operational OPACs. Mandel and Herschman, and Tague, have recommended culling transaction logs to identify search terms that could be added as "see" references to descriptors assigned to the retrieved relevant records.[64] Tague's study "found that the user-supplied keywords retrieved approximately one-third more relevant papers than title keywords alone."[65] She believes that this process is now feasible and would add more current, up-to-date terminology to the system's vocabulary. However, discovering query-term, relevant document associations from transaction logs may require a fair amount of intellectual review. OPACs that ask users to "mark" retrieved items they believe to be relevant to their queries largely resolve this problem. Machine analysis could then "audit" the trail back to the query terms, and either

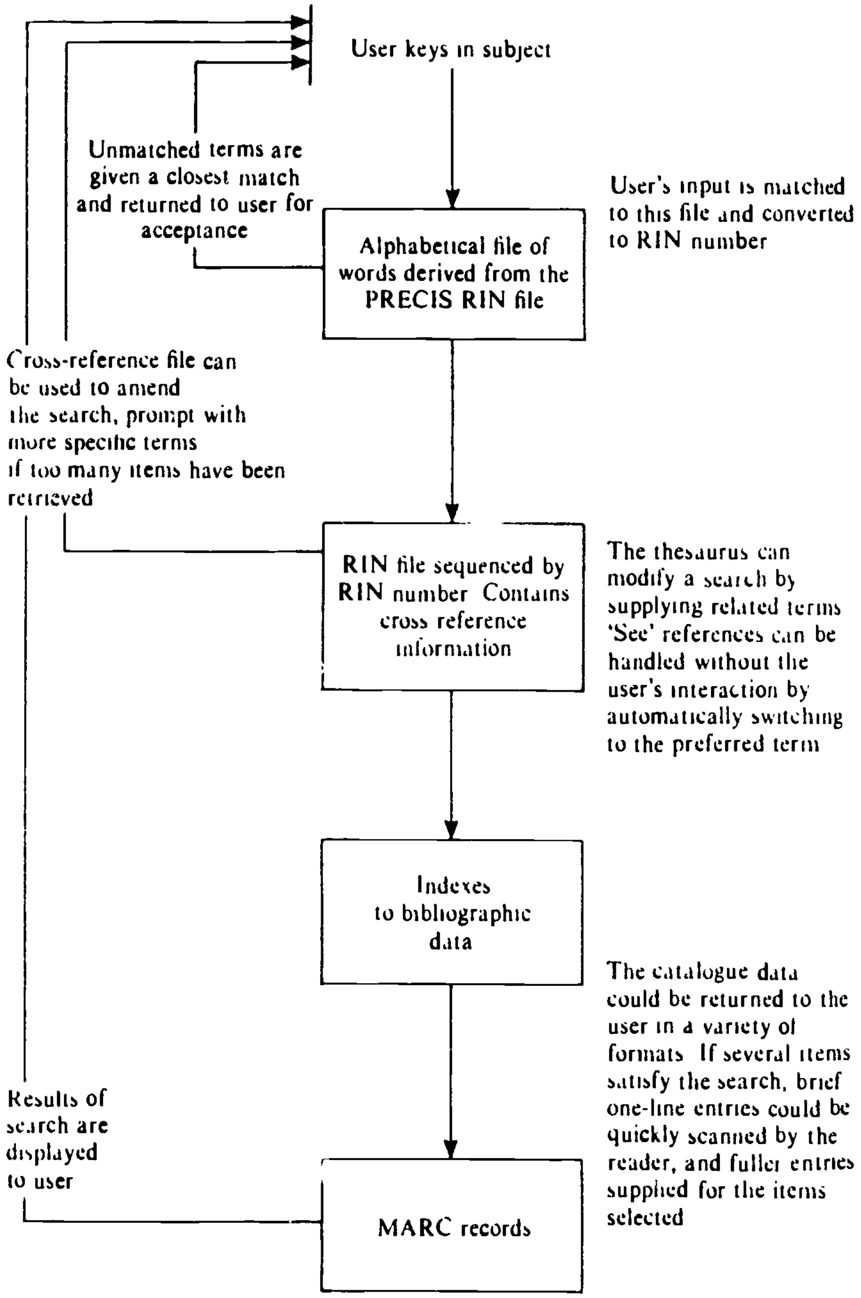

User keys in subject
Unmatched terms are given a closest match and returned to user for acceptance
User's input is matched to this file and converted to RIN number
Alphabetical file of words derived from the PRECIS RIN file
Cross-reference file can be used to amend the search, prompt with more specific terms if too many items have been retrieved
RIN file sequenced by RIN number Contains cross reference information
The thesaurus can modify a search by supplying related terms 'See' references can be handled without the user's interaction by automatically switching to the preferred term
Indexes to bibliographic data
The catalogue data could be returned to the user in a variety of formats If several items satisfy the search, brief one-line entries could be quickly scanned by the reader, and fuller entries supplied for the items selected
Results of search are displayed to user
MARC records

**Figure 30.** *PRECIS OPAC Search Flow, Middlesex Polytechnic, London*

link them automatically to descriptors or produce an output list of candidate "see" references.

## Automatic Natural Language Processing (NLP), Vocabulary Translation and Term-to-Term Mapping

The vocabulary enhancement approach just discussed relies primarily on conventional term matching and document retrieval techniques. Some prototype OPACs and IR system front ends employ automatic language processing techniques to assist with vocabulary control and the translation of search terms to the best or relevant terms in the system's vocabulary. Three such systems are reviewed here: OKAPI, CITE, and ERLI's ALEXIS-based front end.

Automatic natural language processing (NLP) techniques and algorithms have only recently been employed in IR systems and OPACs. The aim of this approach is to support "free-form" natural language expression of information needs and queries by users, and, using automatic mechanisms, to make the translation from this language to relevant system terms. NLP software may also construct the system query (See Figure 22) and trigger the retrieval operation, whatever retrieval method is used. NLP processing often exploits existing system indexes, thesauri, and front-end dictionaries or semantic networks when performing one or more of these functions.

Vocabulary control is needed to alleviate the matching problems caused by the use of natural language in retrieval system queries and indexing. The problems are of three types: morphological, syntactical, and semantic. Thesaurus aids can be used to resolve semantic problems: control of homonyms and synonyms or other term equivalencies, and identification and classification of related, broader, and narrower terms. Most thesauri incorporate both hierarchical and equivalence-type relationships. Users searching under a given term may need to find material indexed under an equivalent term, related terms, or narrower, more precise terms. NLP linguistic analysis techniques have been used successfully in OPACs to assist with morphological (variant but equivalent word forms) and syntactical (variant but equivalent phrases) query-document matching problems. For example, some routines compensate for word spelling or suffix variations; others match direct and inverted forms of subject headings. In some systems these approaches are combined to improve retrieval effectiveness.

OKAPI (See Figure 24) is a prototype third-generation OPAC developed by a research team at the Polytechnic of Central London.[66] It is placed in one of the Polytechnic libraries so actual user sessions can be logged for later analysis. An early version, OKAPI-84, supported natural language subject searching, with weighted-term, combinatoric retrieval. OKAPI-84 also employed search decision tree-based rules to automatically change search strategy when one attempt failed to produce retrieved items. After about two years on site, a random sample of actual searches was selected from the OKAPI-84 log for vocabulary and failure analysis. Although the 1984 version had flexible retrieval routines built in, matching on entry words had to be nearly exact.

Based on this analysis of logged user sessions, OKAPI-86, a newer version, tested several linguistic computing techniques aimed at improving term matching and, thus, to improving search recall: automatic "weak" and "strong" stemming of search terms, automatic cross referencing, and semi-automatic spelling correction. Even though OKAPI-84 employed retrieval methods more flexible than Boolean logic, the researchers were disturbed that subject searches for "accounts dictionary", "terminal illness", and "space shuttle" produced no hits even though the library had books on each subject (the indexed headings are Accounting--dictionaries, Terminal cases, and Reusable space vehicles). Term truncation or being able to browse indexed subject headings might have helped, but the researchers believe a good OPAC should go directly to the records assigned these headings.

The research team installed Porter's stemming algorithms in OKAPI-86.[67] These algorithms were designed to conflate terms that are morphologically similar. The assumption is that they will be semantically close. The "weak" stemming routine normalized singular and plural noun forms, and removed possessive endings, -ed's and -ing's. A stronger routine was tested but later rejected because it often led to a loss of precision in the retrieved results.

OKAPI-86 design also addresses the problem of "words the system cannot find," after the stemming routine has been applied and cross reference lookups have been tried. Most conventional OPACs merely report a failed search if the query contains even one word for which a match cannot be found. OKAPI log analysis revealed that about 10% of searches contained miskeyed or misspelled words. OKAPI-86 automatically invokes a spelling correction procedure which tries to find a correction and then suggests it to the user (See Figure 24). When a word does not stem-match any index entry, the Soundex coding system is used to reduce words for a nearest match in the system's special dictionaries or indexes. If still no match is found, the system ignores the word (a form of implicit stoplisting) and processes the search on any remaining search words.

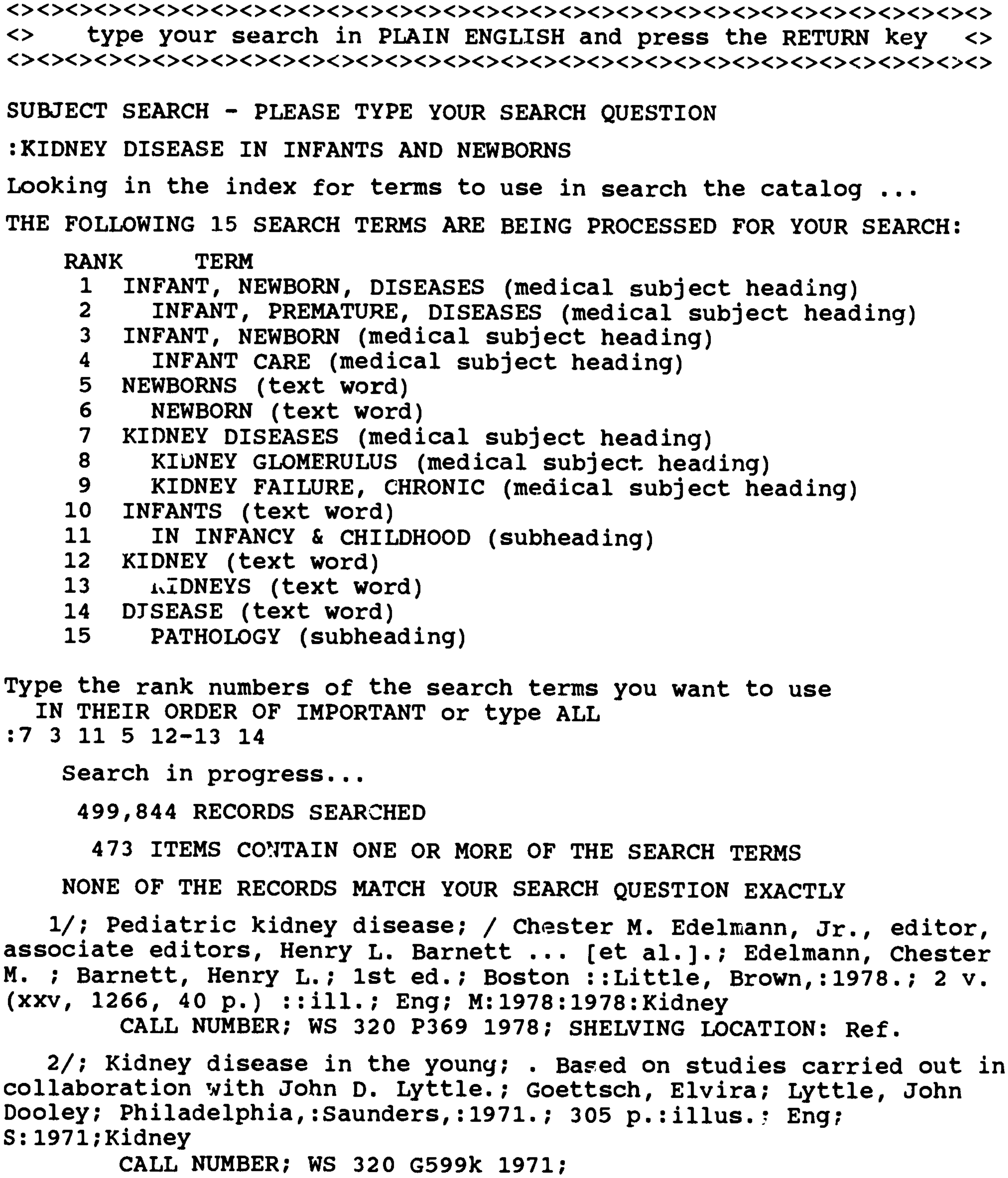
A measure of synonym control was attempted, specifically, automatic cross referencing, using a look-up table which had three kinds of entries: stop words, synonym pairs that would be adversel, affected by conflation (e.g., child, and children), abbreviations and their full expressions words with alternative spellings (jail, gaol), and equivalent word pairs suggested by query terms found in logged searches (Great Britain, UK). The list also included noun phrases to be treated as "words" (shortness of breath, soap opera).

Use of the experimental OKAPI-86 was evaluated carefully through transaction log analysis and user interviews. In some cases users' searches were repeated for comparative purposes by the researchers on another OKAPI OPAC which did not employ these linguistic analysis and processing mechanisms. The weak stemming procedure was shown to increase recall in most searches with no corresponding significant increase in non-relevant items retrieved. The experimental OKAPI-86 OPAC corrected about half of the spelling/miskeying errors with favorable results. The cross referencing always helped when it was called into play. About 25% of the searches studied contained a word or phrase that matched one in the cross references list. This was not a surprise to the research team because the list was largely constructed from search terms that had been input in the past by users of the same library. In almost all of the cases of an automatic reference switch, recall was increased without a decrease in precision. Very clearly, OKAPI-86 has shown that OPAC subject retrieval performance can be improved through the use of automatic linguistic term matching aids.

CITE, the third-generation OPAC at the National Library of Medicine (NLM), supports natural language queries and performs intelligent stemming on the user's search terms. Stemmed query words are looked up in both free text indexes and the MeSH (Medical Subject Headings) thesaurus. Search words found in free text are then automatically linked to associated MeSH descriptors. The CITE retrieval methods include term weighting, combinatoric searching, relevance feedback, query expansion, and ranked output. In the combinatoric, term weighting retrieval method, matches on all combinations of the translated query terms are retrieved and ranked for output in decreasing order of relevance. Documents estimated to be most relevant are output at the top of the list.

To begin a subject search, CITE invites the user to "TYPE YOUR SEARCH QUESTION" (See Figure 31).[68] Significant words in the query (those not stoplisted) are passed through a customized stemming algorithm which conflates variant word forms, but also removes endings very common in medical vocabulary (e.g., -itis, -ectomy). Rule-based stemming algorithms are a rudimentary form of "expert" systems software. The "accepted" stems are then matched against the text and controlled vocabularies of the system. All search terms that have been selected as "good" are then displayed to the user for personal selection and ranking. The initial indirect semantic mapping to thesaurus terms, which is a form of synonym/related term control, may continue if the user chooses to see items related to a retrieved item he has judged as relevant. MeSH headings assigned to the relevant documents are used in the extended search. Even without employing semantic aids such as synonym tables, or a front-end dictionary or semantic network which would map potential search terms with related MeSH descriptors, CITE "achieves a considerable degree of semantic query expansion" using its automatic and semi-automatic processes.[69] To this author's knowledge, unlike OKAPI and the ERLI software, CITE does not perform any syntactical analysis or processing of query terms to identify/normalize, for example, compound noun phrases.

The Paris firm FRLI (Etude et Recherche en Linguistique et Informatique) has developed natural language database access software which uses the firm's proprietary ALEXIS database management software. This complex AI-based natural language retrieval front end package can only be summarized here. ERLI's intelligent front end (sometimes referred to as ALEX-DOC, but here labelled



```
<><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><>
<>   type your search in PLAIN ENGLISH and press the RETURN key  <>
<><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><>

SUBJECT SEARCH - PLEASE TYPE YOUR SEARCH QUESTION

:KIDNEY DISEASE IN INFANTS AND NEWBORNS

Looking in the index for terms to use in search the catalog ...

THE FOLLOWING 15 SEARCH TERMS ARE BEING PROCESSED FOR YOUR SEARCH:

    RANK     TERM
     1  INFANT, NEWBORN, DISEASES (medical subject heading)
     2    INFANT, PREMATURE, DISEASES (medical subject heading)
     3  INFANT, NEWBORN (medical subject heading)
     4    INFANT CARE (medical subject heading)
     5  NEWBORNS (text word)
     6    NEWBORN (text word)
     7  KIDNEY DISEASES (medical subject heading)
     8    KIDNEY GLOMERULUS (medical subject heading)
     9    KIDNEY FAILURE, CHRONIC (medical subject heading)
    10  INFANTS (text word)
    11    IN INFANCY & CHILDHOOD (subheading)
    12  KIDNEY (text word)
    13    KIDNEYS (text word)
    14  DISEASE (text word)
    15    PATHOLOGY (subheading)

Type the rank numbers of the search terms you want to use
  IN THEIR ORDER OF IMPORTANT or type ALL
:7 3 11 5 12-13 14

    Search in progress...

     499,844 RECORDS SEARCHED

      473 ITEMS CONTAIN ONE OR MORE OF THE SEARCH TERMS

    NONE OF THE RECORDS MATCH YOUR SEARCH QUESTION EXACTLY

   1/; Pediatric kidney disease; / Chester M. Edelmann, Jr., editor,
associate editors, Henry L. Barnett ... [et al.].; Edelmann, Chester
M. ; Barnett, Henry L.; 1st ed.; Boston ::Little, Brown,:1978.; 2 v.
(xxv, 1266, 40 p.) ::ill.; Eng; M:1978:1978:Kidney
        CALL NUMBER; WS 320 P369 1978; SHELVING LOCATION: Ref.

   2/; Kidney disease in the young; . Based on studies carried out in
collaboration with John D. Lyttle.; Goettsch, Elvira; Lyttle, John
Dooley; Philadelphia,:Saunders,:1971.; 305 p.:illus.; Eng;
S:1971;Kidney
        CALL NUMBER; WS 320 G599k 1971;
```

**Figure 31.** *CITE*

"NLQP" for natural language query processor) can be adapted to a variety of retrieval/database systems, including OPACs. It may be interfaced to existing IR systems, or it can manage both a thesaurus and document retrieval system.

ERLI has developed NLQPs for interactive consultation of the "yellow pages in the French electronic telephone directory and for lookups in the directory of Teletel services available in France through the Minitel terminals placed in millions of French homes and businesses. The effectiveness of the ERLI NLQP for bibliographic retrieval is currently being tested. It has been adopted as the retrieval software for access to the French online subject authorities file, called RAMEAU. This file is maintained by the Bibliotheque Nationale and the Centre National du Catalogue Collectif National (CCN), and it is available as a "public" file to libraries via SUNIST, the national university network for scientific and technical information. This authority file and the retrieval software reside on a SUNIST host IBM computer near the city of Lyon.

Of particular interest here are the advanced, AI-based linguistic processing methods used by ERLI's expert query system to effect natural language query matching and retrieval in large thesaurus-controlled databases. ERLI views the IR process as an "intellectual translation operation" which consists of three basic tasks:

- relating the terms of a user's question to those of the documentary language used,

- deciding which are the terms of the documentary language best corresponding to what the user is seeking,

- editing a Boolean search equation, with the terms retained, which most closely translates the user's question.[70]

Based on an analysis of the knowledge and intellectual operations used by expert search intermediaries, ERLI developed NLQP as an expert system capable of performing these functions automatically, while permitting the user to enter queries in natural language and to interactively modify the system's query formulation. NLQP also permits the user to browse in the online thesaurus at the initial stage of the retrieval process. Figure 32 illustrates with a block diagram the way in which NLQP works.

To support the performance of its automated expert assistance functions, NLQP relies on a knowledge base consisting of a thesaurus, an associated dictionary, and a rich set of rules for linguistic analysis and transformation. The thesaurus may be any structured (hierarchical or network), controlled vocabulary of subject descriptors or headings. The dictionary is a special- purpose lexicon with links connecting terms in the lexicon and these terms with those in the thesaurus. These links model lexical, syntactical and semantic knowledge, the role of which is to insure the passage from natural language to that of the biblio-

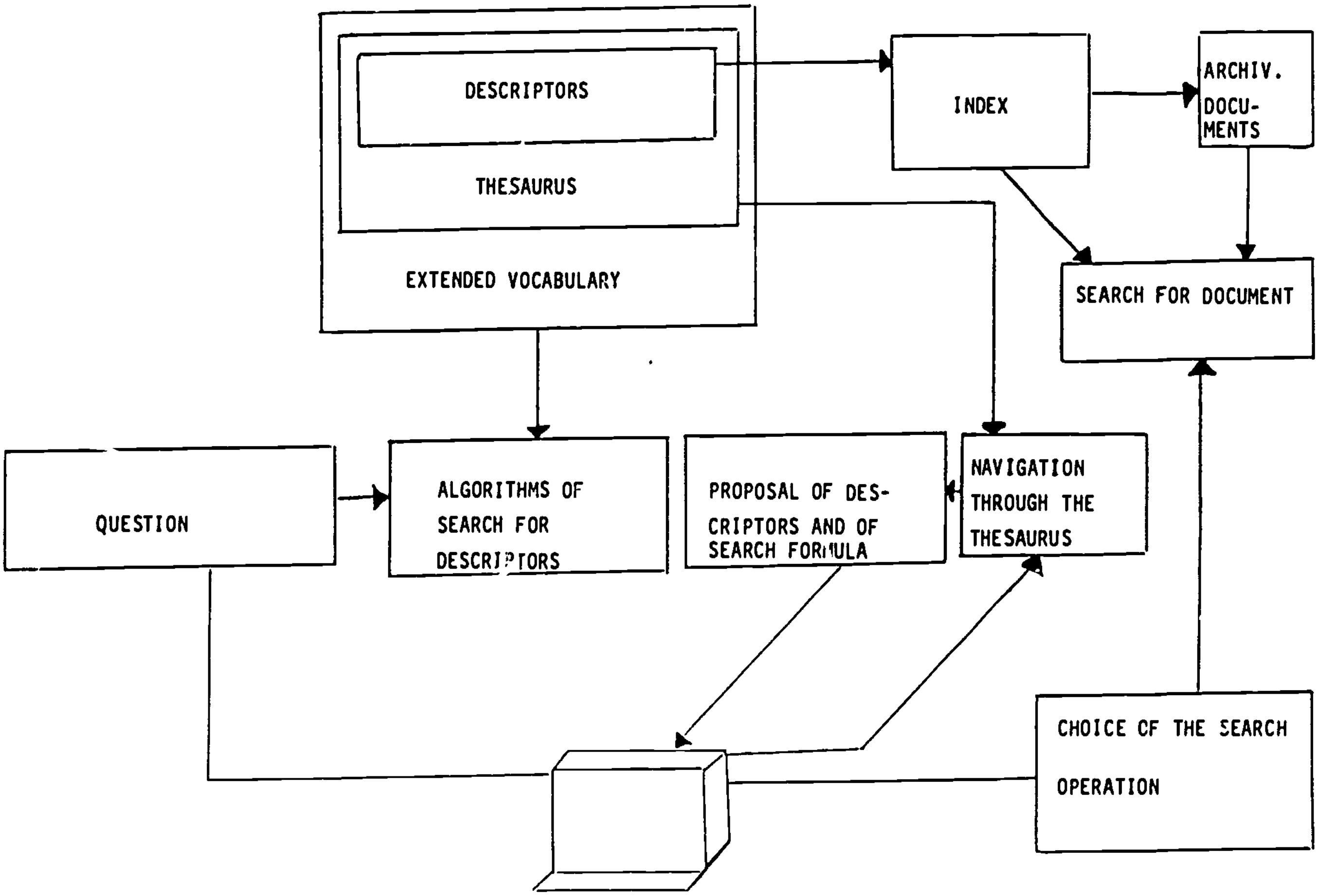

DESCRIPTORS
THESAURUS
EXTENDED VOCABULARY
INDEX
ARCHIV. DOCU-MENTS
SEARCH FOR DOCUMENT
QUESTION
ALGORITHMS OF SEARCH FOR DESCRIPTORS
PROPOSAL OF DES-CRIPTORS AND OF SEARCH FORMULA
NAVIGATION THROUGH THE THESAURUS
CHOICE OF THE SEARCH OPERATION

**Figure 32.** *Diagram of ERLI's Expert Query Processor*

graphical database. The rule set models the algorithmic processing required to transform natural language requests into descriptors in the thesaurus. They specify how a given linguistic construct is to be processed in a given situation.

Search requests may be entered in natural language: for example, "safety controls in nuclear reactors" or "space shuttle engineering faults." No command language is required, and no effort on the part of the user to match the system's vocabulary is necessary. The rule-governed NLQP module's attempt to find the best descriptors for a query may be performed in up to four phases. Each phase may activate a series of analysis/transformation steps. The phases are defined as follows:

1. General text processing: sentence demarcation;

2. Morphological and lexical analysis:
   - Recognition of words,
   - Recognition of expressions (i.e., compound units),
   - Removing of ambiguities ("disambiguisation");

3. Syntactical analysis:
   - Phrase demarcation (phrases are syntactic functional units),
   - Matching with syntactic structures (phrase restructuration);

4. Semantical analysis:
   - Lexical substitution,
   - Proposition of descriptors.

During the first phase of NLQP, rudimentary morphological and phrase analysis takes place to demarcate words and/or sentences and to edit out insignificant typological characters. A second phase of morphological and lexical analysis identifies common word forms, compound word expressions, and checks for their existence in the thesaurus and its associated dictionary. Proprietary stemming and spelling correction algorithms are applied at this point to help find the word in the dictionary. A third phase of syntactical analysis recognizes and operates on compound word phrases such as noun phrases, conjunctive expressions, and prepositional phrases. The analysis at this stage attempts to "understand" the meaning of the phrase through its structure for subsequent transformation to the correct descriptor, which may have a very different syntactical makeup: for example, user query terms like "presidential election" or "primary elections" transformed to "Election of the President."

The fourth phase consists of a flexible, structured, combinatoric dictionary/thesaurus look-up process. Matches are attempted against a variety of components or reconstructions of the original query expression. These decomposition, restructuring rules "know" all the syntactical structures used in the given thesaurus. Transformation of phrases may take place on the basis of word addition or omission, word coordination or apposition, and exchange of a noun for an adjective (or vice versa). Finally (perhaps, this is often an iterative process), each component word of a phrase is stemmed and may be replaced by its equivalent word listed in the thesaurus or its synonym dictionary.

To summarize, the NLQP work is carried out at three levels of analysis: stemming and spelling correction operations on words; structural analysis of compound terms until a match is found or the rules are exhausted; and use of the thesaurus to identify additional related, broader, or narrower terms. A fifth phase involves the automatic formulation of a Boolean query using the descriptors identified in the language ~ ılysis steps outlined above.

## 5.3 The Contextual Subject Approach: Improved Browsing and Related-Item Navigation Facilities

Finding the best term or terms for a relevant match and, perhaps, identifying related terms that can be used to discriminate and focus a search or to expand it: that is the non-trivial task facing the user at the search interface. Information scientists have been trying to gain an understanding of this need and its requirements. Tague rejects one conventional interpretation:

> What, then, is this process going on when a user stands in front of an OPAC terminal? It is something more than just a matching or translation, and the reason it is more is that the user and the system do not understand each other perfectly.[71]

Tague argues that the interface process is, or should be, a negotiation process: "Matching and translation imply a one-way process." The negotiation or bargaining which takes place "is over the language which must be used in order to extract the most useful output from the store of bibliographic items." The user is seeking information about the way the system describes its materials. The system should be seeking information about the way the user describes wanted materials, or at least be seeking indicators of the user's problem or need. Interactive dialogue is the means of affecting this negotiation.

But does this say anything more than, "Unfriendly nations should talk to each another."? The policies, protocols, and procedures must be developed in detail. Speaking the same language, of course, helps; but this is not always possible. A demonstrated, displayed spirit of cooperation is, however, required. Now, to the details.

Words, and their meanings, are at the heart of the problem. Subject *headings* or *descriptors* (these words are synonymous here) typically contain more than one word. Hc··ver, most subject searches entered in OPACs contain only one or two words. The single word looms large in current matching techniques.

In phrase-match searching, everything depends on matching that first word in the heading. And, of course, keyword searching matches word to word, whatever Boolean computations follow.

One word may have many different meanings or shadings of meaning (the homonym problem). Ignore, for the moment, variant spellings of the same word (orthography) and variant forms of the same word (morphology). Even if one believes a word like "forecasting" has a general or natural definition, there is no denying it acquires different meanings *in use* in different subject fields or contexts (e.g., public finance, business, meteorology, horse racing). In fact, the meaning of a word or a phrase may best be grasped in its use, in this context or that.

In a narrow, "well-defined" subject field or domain it may be possible to predict the meaning-in-use of a word. This predictability greatly simplifies the matching/translation problem. However, most library catalogs describe and index materials in many, many different subject areas, each likely to include many different aspects treated from a variety of perspectives and levels. Most existing thesauri are specific to a subject or discipline. Matching a query word with a word in one of their headings gets one pretty close to relevant materials. LCSH and PRECIS, for example, are not subject-specific. Their headings cover the whole lot of published materials. Simply matching a word or two in these headings may not even get you into the ballpark. Bates hopes for OPACs which will begin to work if the user merely hits the "side of the barn." What if it is the wrong barn?

Thus, subject searching in OPACs requires the "contextual subject approach" if a large part of the vocabulary problem is to be resolved. There are a number of ways to implement this to good effect. A two-way process is best: the system can do some things better, or at least faster, than the typical user (e.g., spelling corrections, dictionary lookups, phrase restructuring), and only the user can ultimately decide the issues of "best" term and document relevance. Also, as we have seen, delegating certain functions to the system permits users to express their queries in natural language. "Negotiation" has a negative ring about it. "User engagement" is preferred.

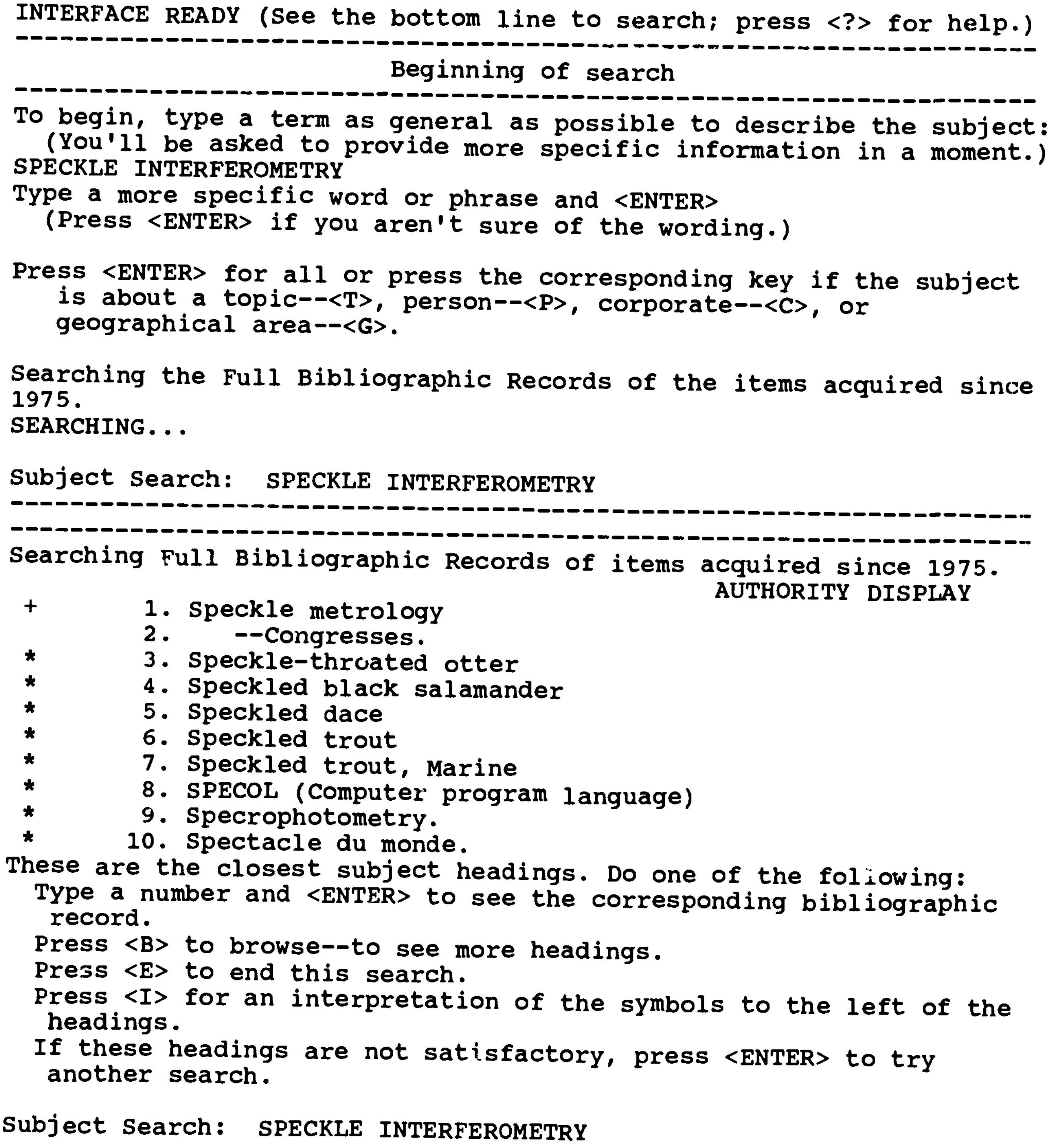
What dialogue, database, and display design features can help? Present-day OPACs offer little in the way of vocabulary contextual aids. By design or not, the following features in today's OPACs provide some limited assistance: alphabetically-ordered term browsing lists, plain English field labels in displayed citations (e.g., "SUBJECT:"), and the highlighting of words or phrases in the displayed citation that matched search terms and brought about the retrieval. The last two features reveal to the user the matching subject word's or heading's use in some context. The first feature, a displayed list of subject terms or headings merely indicates that the terms or headings are used in the database. Qualifying words or subheadings attached to the main descriptor may give some indication of its contextual meaning. For the most part, current OPAC alphabetically-arranged subject term browsing lists explain neither the meaning of individual LCSH descriptors nor their role in the database.

Putting LCSH online in its present form for consultation by end users is not the answer. Substantive LCSH limitations aside, users are not likely to consult LCSH prior to their initial searching activities. How many studies do we need to convince us that the "red books" placed near the catalog are seldom consulted by catalog users? A thesaurus is a tool for an expert. As Shoval notes: "A user of a retrieval system may neither be able nor be willing to spend time and effort to study how to use it."[72] In addition, in our daily reading and discourse, be it speech or writing, how many of us pick up a dictionary right off?

OPAC users want displays of related terms and intelligent, understandable online vocabulary guidance. With a better understanding of the motivational and cognitive factors involved, system designers are providing new solutions for OPAC subject searchers based on the contextual subject approach. These solutions include: 1) explicit dialogue linkages between query terms and the controlled vocabulary in retrieved citations, 2) the use of retrieved, displayed citations as a springboard or gateway to related materials, 3) the use of citations judged by the searcher to be relevant as a source of vocabulary for automatic query expansion, and 4) the innovative use of new windowing and graphics software (and large, high-resolution VDUs) to display thesaural term-term and term-document relationships.

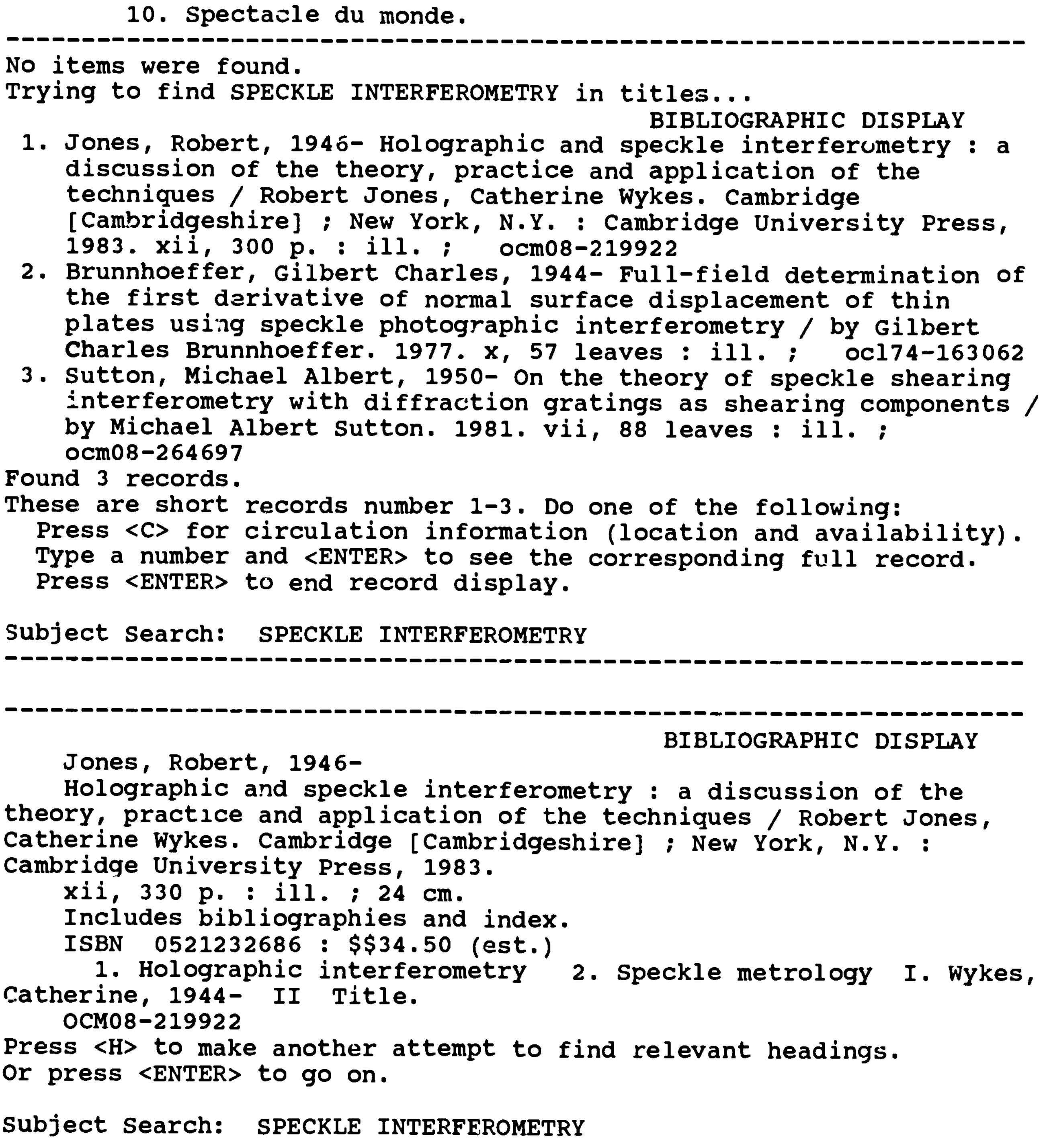
The OPAC at the University of Illinois at Urbana-Champaign (UIUC) is comprised of the LCS short-record circulation system and the full-MARC record WLN search system. IBM PC microcomputers are employed as public user search terminals. Specially designed user interface software which runs on each PC terminal simplifies search entry and includes intelligent search and vocabulary aids to improve subject retrieval in the OPAC.

Figure 33 contains a series of six screens to illustrate some of these features. After selecting "Subject" from the main search menu, the user is prompted to enter a "general" term first, followed by a request from the system for a "specific word or phrase." The option list to permit further narrowing of the search is then displayed. If the subject search statement does not match with any entry in the subject headings index (a "nearby" portion is displayed), the system proceeds: "Trying to find SPECKLE INTERFEROMETRY in titles." This leads to an automatic keyword-in-title, Boolean AND search. If the search produces results, brief citations of matching records are displayed and the user is permitted to request a full citation display for any one of them. If a full citation is displayed, a prompt to "Press <H> to make another attempt to find relevant headings" appears. If this prompt is followed, the interface software automatically selects and displays an associated subject heading and any linked cross references. If the searcher chooses to go forward at this point, the system retrieves records indexed under the system-selected subject heading.



```
INTERFACE READY (See the bottom line to search; press <?> for help.)
--------------------------------------------------------------------
                           Beginning of search
--------------------------------------------------------------------
To begin, type a term as general as possible to describe the subject:
  (You'll be asked to provide more specific information in a moment.)
SPECKLE INTERFEROMETRY
Type a more specific word or phrase and <ENTER>
  (Press <ENTER> if you aren't sure of the wording.)

Press <ENTER> for all or press the corresponding key if the subject
   is about a topic--<T>, person--<P>, corporate--<C>, or
   geographical area--<G>.

Searching the Full Bibliographic Records of the items acquired since
1975.
SEARCHING...

Subject Search:  SPECKLE INTERFEROMETRY
--------------------------------------------------------------------
--------------------------------------------------------------------
Searching Full Bibliographic Records of items acquired since 1975.
                                                 AUTHORITY DISPLAY
 +       1. Speckle metrology
         2.    --Congresses.
 *       3. Speckle-throated otter
 *       4. Speckled black salamander
 *       5. Speckled dace
 *       6. Speckled trout
 *       7. Speckled trout, Marine
 *       8. SPECOL (Computer program language)
 *       9. Specrophotometry.
 *      10. Spectacle du monde.
These are the closest subject headings. Do one of the following:
  Type a number and <ENTER> to see the corresponding bibliographic
   record.
  Press <B> to browse--to see more headings.
  Press <E> to end this search.
  Press <I> for an interpretation of the symbols to the left of the
   headings.
  If these headings are not satisfactory, press <ENTER> to try
   another search.

Subject Search:  SPECKLE INTERFEROMETRY
```

**Figure 33.** *University of Illinois at Urbana-Champaign OPAC*



```
        10. Spectacle du monde.
-----------------------------------------------------------------------
No items were found.
Trying to find SPECKLE INTERFEROMETRY in titles...
                                               BIBLIOGRAPHIC DISPLAY
 1. Jones, Robert, 1946- Holographic and speckle interferometry : a
    discussion of the theory, practice and application of the
    techniques / Robert Jones, Catherine Wykes. Cambridge
    [Cambridgeshire] ; New York, N.Y. : Cambridge University Press,
    1983. xii, 300 p. : ill. ;   ocm08-219922
 2. Brunnhoeffer, Gilbert Charles, 1944- Full-field determination of
    the first derivative of normal surface displacement of thin
    plates using speckle photographic interferometry / by Gilbert
    Charles Brunnhoeffer. 1977. x, 57 leaves : ill. ;   ocl74-163062
 3. Sutton, Michael Albert, 1950- On the theory of speckle shearing
    interferometry with diffraction gratings as shearing components /
    by Michael Albert Sutton. 1981. vii, 88 leaves : ill. ;
    ocm08-264697
Found 3 records.
These are short records number 1-3. Do one of the following:
  Press <C> for circulation information (location and availability).
  Type a number and <ENTER> to see the corresponding full record.
  Press <ENTER> to end record display.

Subject Search:  SPECKLE INTERFEROMETRY
-----------------------------------------------------------------------

-----------------------------------------------------------------------
                                                BIBLIOGRAPHIC DISPLAY
    Jones, Robert, 1946-
    Holographic and speckle interferometry : a discussion of the
theory, practice and application of the techniques / Robert Jones,
Catherine Wykes. Cambridge [Cambridgeshire] ; New York, N.Y. :
Cambridge University Press, 1983.
    xii, 330 p. : ill. ; 24 cm.
    Includes bibliographies and index.
    ISBN  0521232686 : $$34.50 (est.)
      1. Holographic interferometry   2. Speckle metrology  I. Wykes,
Catherine, 1944-  II  Title.
    OCM08-219922
Press <H> to make another attempt to find relevant headings.
Or press <ENTER> to go on.

Subject Search:  SPECKLE INTERFEROMETRY
```

**Figure 33.** *University of Illinois at Urbana-Champaign OPAC (Continued)*

```
                                                BIBLIOGRAPHIC DISPLAY
    Jones, Robert, 1946-
    Holcgraphic and speckle interferometry : a discussion of the
theory, practice and application of the techniques / Robert Jones,
Catherine Wykes. Cambridge [Cambridgeshire] ; New York, N.Y. :
Cambridge University Press, 1983.
    xii, 330 p. : ill. ; 24 cm.
    Includes bibliographies and index.
    ISBN  0521232686 : $$34.50 (est.)
      1. Holographic interferometry   2. Speckle metrology  I. Wykes,
Catherine, 1944-  II. Title.
    OCM08-219922
A relevant heading is:  Holographic interferometry
                                                   AUTHORITY DISPLAY
   Holographic interferometry
    SA  Holography
Later on you may wish to search the SA (See Also) heading(s) above.
Press <ENTER> to go on.

Subject Search:  SPECKLE INTERFEROMETRY
-----------------------------------------------------------------
-----------------------------------------------------------------
A relevant heading is:  Holographic interferometry
                                                   AUTHORITY DISPLAY
   Holographic interferometry
    SA  Holography
Heading:   Holographic interferometry
                                                BIBLIOGRAPHIC DISPLAY
  1. Jones, Robert, Holographic and speckle interferometry : a
     discussion of the theory, practice and application of the
     techniques / 1983. xii, 300 p. ocm08-219922
  2. Weimer, David. Pockels-effect call for gas-flow simulation /
     1982.   OCM08-441154
  3. Schumann, Walter. Holographic interferometry : from the scope of
     deformation analysis of opaque bodies / 1979. 194 p. ; 24 cm.
     ocl74-804311
  4. Ostrovsky, Yu. I. Interferometry by holography / 1980.  x, 330 p.
     ocl75-893949
  5. Schu... nn, W. Holography and deformation analysis / c1985. x, 234
     p.    ocm12-134876
Found 5 records.
These are short records number 1-5. Do one of the following:
   Press <C> for circulation information (location and availability).
   Type a number and <ENTER> to see the corresponding full record.
   Press <ENTER> to end record display.

Subject Search:  Holographic interferometry
```

**Figure 33.** *University of Illinois at Urbana-Champaign OPAC (Continued)*

There are two or three notable features incorporated into this smart user interface. Phrase match subject searches are not permitted to fail; they are automatically reformulated into title keyword searches. The software guides the searcher via the dialogue from uncontrolled title words to the subject heading(s) assigned to works in which the title words appear. When available, subject heading "see also" references are also displayed. If the user desires, the software performs a new subject search on the heading it has singled out. Throughout this interaction the user is engaged in a nice mix of delegated and augmented assistance from the interface software.

NLM's CITE OPAC (Figure 21) software runs on the computer which hosts the database management and basic retrieval software. The CITE interface software translates the user's subject query and returns a display of suggested searching vocabulary which includes both free text terms and associated MeSH subject headings. The user may select and/or rank terms he wishes to include in the search. "Closest-match" citations are then displayed in ranked order. After viewing a few, the searcher may choose to see "items related" to a citation of interest. The system then reports to the user the new terms and descriptors it has selected to conduct the follow-on search for related items. The new search terms are derived primarily from the bibliographic record selected by the user as relevant or interesting. The user may add to, select or re-rank terms on this new list, and so on.

Simple dialogue messaging guides the CITE user through these stages, but the "engaged" user makes the major control decisions. Both free text terms and associated controlled vocabulary terms are displayed together, and retrieved records judged relevant by the searcher are used as a source of descriptors for expanded searching down potentially relevant pathways. Paperchase, the OPAC/IR system pioneered at the Med'cal Library of Boston's Beth Israel Hospital, matches words from search requests to document titles and immediately displays a list of MeSH descriptors assigned to works having those words in their titles.[73]

TINlib, an OPAC based on Information Management & Engineering (IME) Limited's TINMAN database management and retrieval software, employs a navigation feature that can be customized to establish almost any search pathways through a database. Based on the flexible database modelling and data linking facilities supported under its "entity-relational" database architecture, TINMAN permits the precoordination of search linkages from any field (an entity) to any other field. Several different search and navigation pathways can be designed into a single TINlib OPAC. Users may choose a traditional search mode (e.g., browse or keyword queries), but are permitted also to point to a data element in a displayed citation and navigate (or "jump") to related terms and/or related documents in one or two "press" steps. (See Figure 29). In the example shown, the user is taken to a mixed thesaurus display. Any thesaurus or term-term relationships included and modelled in the original OPAC database design can be displayed at this stage. The jumping-off point may include any data element in the citation (e.g., publisher, subject, class number) because data elements rather than catalog records are viewed as the basic units or entities in the database, and any one entity type can be linked to any other. Perhaps odd, but someone might wish to link authors and publishers. The links are two ways; one might identify an author in a search, point to the displayed publisher, and navigate to a list of authors whose works have been published by that publisher. Such entity links can be established both within and among thesauri.

TINMAN/TINlib permits subject browsing at the citation level. Subject clues or pointers in a displayed citation may lead directly down a path to related documents. Pointing to a subject heading in a citation may lead the searcher into the thesaurus to review related terms associated with the document considered relevant. Movement from the thesaurus display to linked citations requires only a "point" and a "press" from the user.

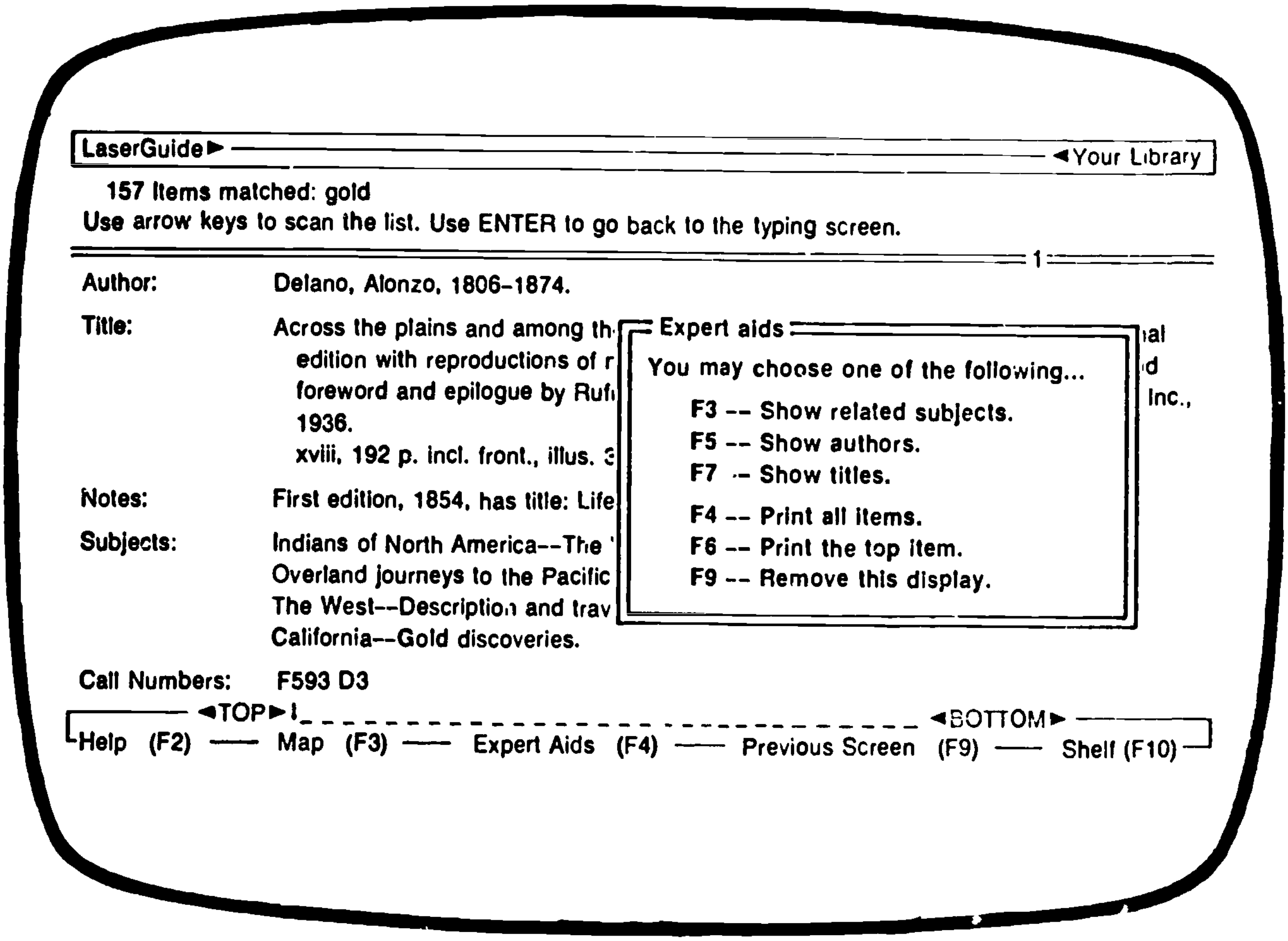
Peter Willett, designer of INSTRUCT, a demonstration and training IIR system developed at the University of Sheffield (UK), describes a mode of navigation like TINlib's as "chain searching."[74] In a document search, after identifying a relevant document, the searcher can invoke a chain search which leads down trails of related documents. INSTRUCT uses statistical "nearest neighbor" routines to collocate the related documents. TINlib's related document linkages must be predefined as part of the early database modelling and mapping activities. This is a profiling operation that does not require programming.

INSTRUCT also allows a document to be used for "seed searching." Seed searching invokes an expanded-query, "best match" search process in which the user's original query words (stemmed) are replaced by the word stems in the title and abstract of a document which the user has judged to be relevant.

> The advantages of this approach are that the user can follow up a particularly interesting document, without losing the main thread of the query, and can move away from the original query to wander in and out of different areas of interest, developing the query as they go along. *Such facilities are particularly well-suited to situations where users are unfamiliar with the scope or contents of a database or where they are unable to specify a precise information need.*[75] (my emphasis)

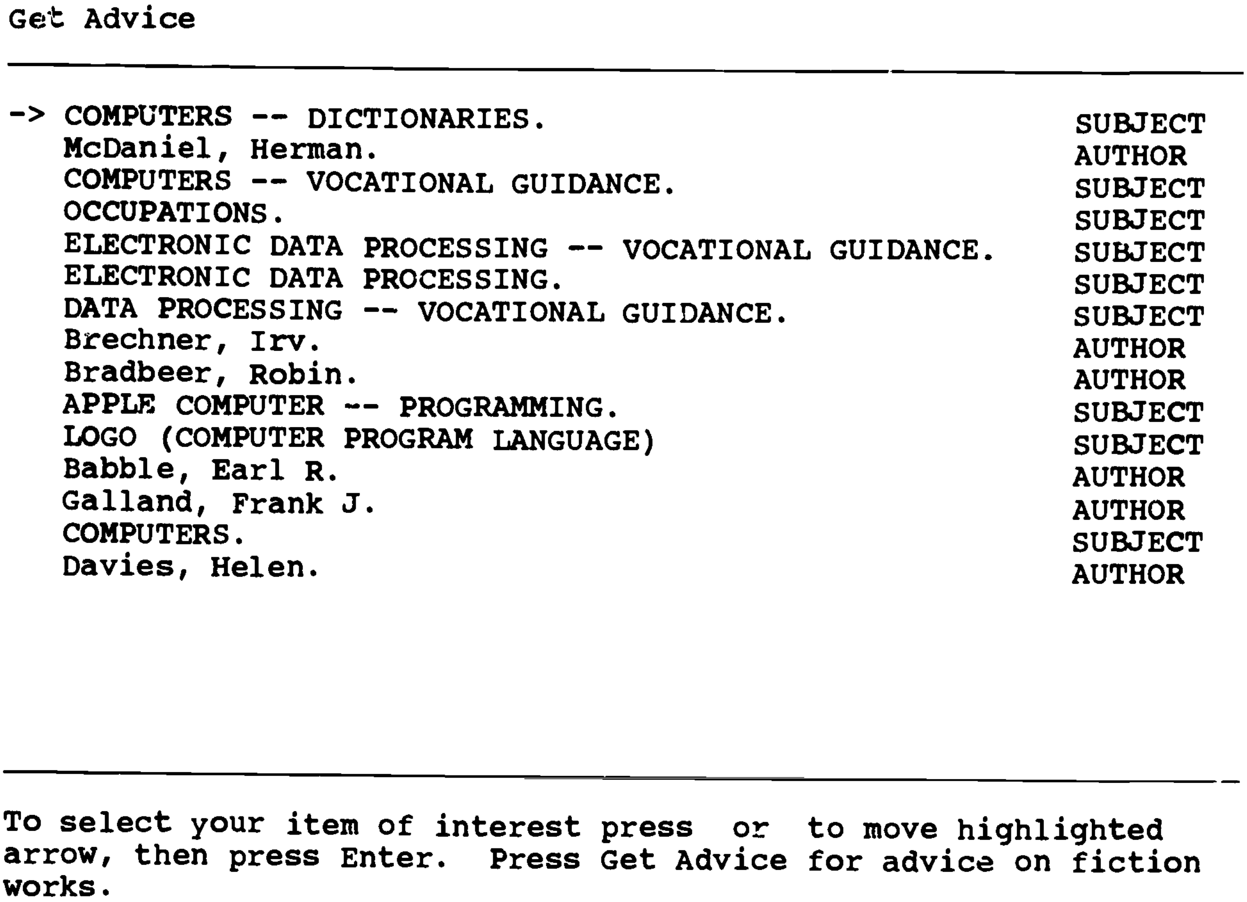
General Research Corporation's (GRC) "LaserGuide" CD-ROM OPAC has a related-subject search feature similar to TINlib's navigation feature (See Figure 34). From a displayed citation, the user may request the "Expert Aids" option. A menu of options appears in a small window box. The choices include "Show related subjects." The press of a function key invokes the option selected.

BiblioFile's "Intelligent Catalog" goes a little further. The system keeps track of a current search. If the searcher is frustrated or not satisfied with results at a stage of the search, he can "Get Advice" as a new option. (See Figure 35). This function has the system suggest other related headings that may



```
LaserGuide►                                                  ◄Your Library
  157 Items matched: gold
 Use arrow keys to scan the list. Use ENTER to go back to the typing screen.
                                                                 1
 Author:        Delano, Alonzo, 1806-1874.
 Title:         Across the plains and among th  Expert aids                    ial
                  edition with reproductions of r  You may choose one of the following...   d
                  foreword and epilogue by Rufi      F3 -- Show related subjects.          Inc.,
                  1936.                               F5 -- Show authors.
                  xviii, 192 p. incl. front., illus. 3  F7 -- Show titles.
 Notes:         First edition, 1854, has title: Life   F4 -- Print all items.
 Subjects:      Indians of North America--The '        F6 -- Print the top item.
                Overland journeys to the Pacific        F9 -- Remove this display.
                The West--Description and trav
                California--Gold discoveries.
 Call Numbers:  F593 D3
        ◄TOP► 1                                              ◄BOTTOM►
 Help (F2) —— Map (F3) —— Expert Aids (F4) —— Previous Screen (F9) —— Shelf (F10)
```

**Figure 34.** *LaserGuide Expert Aids Menu*



```
Get Advice
_____________________________________________________________________________

-> COMPUTERS -- DICTIONARIES.                                          SUBJECT
   McDaniel, Herman.                                                   AUTHOR
   COMPUTERS -- VOCATIONAL GUIDANCE.                                   SUBJECT
   OCCUPATIONS.                                                        SUBJECT
   ELECTRONIC DATA PROCESSING -- VOCATIONAL GUIDANCE.                  SUBJECT
   ELECTRONIC DATA PROCESSING.                                         SUBJECT
   DATA PROCESSING -- VOCATIONAL GUIDANCE.                             SUBJECT
   Brechner, Irv.                                                      AUTHOR
   Bradbeer, Robin.                                                    AUTHOR
   APPLE COMPUTER -- PROGRAMMING.                                      SUBJECT
   LOGO (COMPUTER PROGRAM LANGUAGE)                                    SUBJECT
   Babble, Earl R.                                                     AUTHOR
   Galland, Frank J.                                                   AUTHOR
   COMPUTERS.                                                          SUBJECT
   Davies, Helen.                                                      AUTHOR


_____________________________________________________________________________
To select your item of interest press  or  to move highlighted
arrow, then press Enter.  Press Get Advice for advice on fiction
works.
```

**Figure 35.** *BiblioFile "Get Advice" Feature*

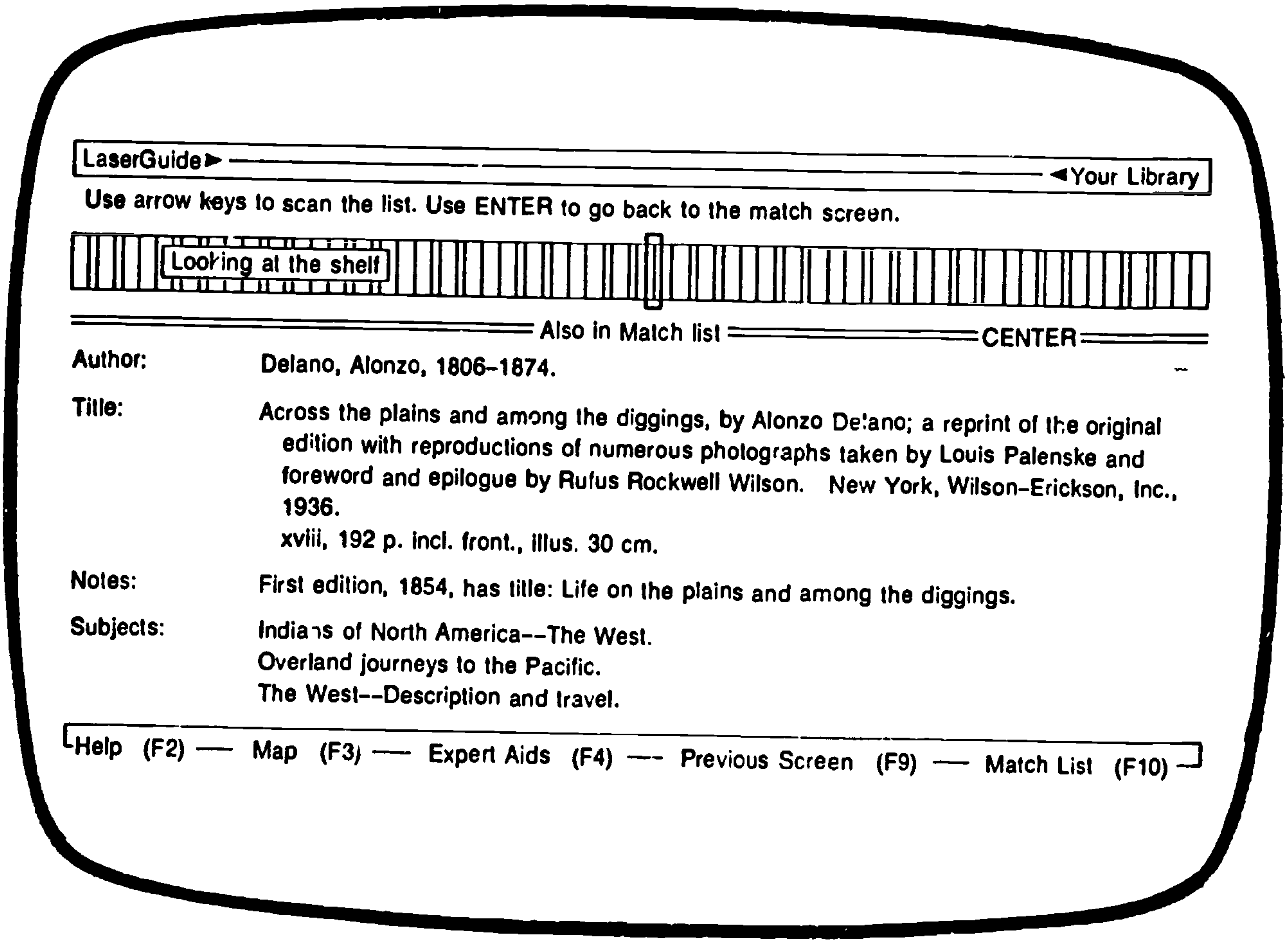
be searched for new or additional items of interest. Tracking a searcher's progress (or lack of it) is a fundamental requirement of intelligent systems software that is able to intervene automatically or semi-automatically to assist the searcher or adapt to the searcher's needs of the moment.

Both the GRC and BiblioFile OPACs provide very sensible shelf list browsing facilities. From a displayed citation in a classed subject area, whether or not the citation is relevant, the searcher may select a full citation display of the item shelved to its left or right simply by using the arrow keys (See Figure 36). The on-screen citation changes accordingly as the arrows are moved to the left or right. No intermediate scanning through single- line uninformative listings is required. This feature moves OPACs closer to what is known in computing as "direct manipulation" or "object-oriented" interfaces. The objects desired (e.g., book records - How nice table of contents would be here!) are manipulated in a direct, more intuitive manner, avoiding previous layers of mental encoding/decoding and indirect representation searchers are usually required to pass through. BiblioFile's "Browse Topics" search option (See Figure 25) leads the searcher to this shelf browsing facility through a series of subject table displays drawn from the Dewey or LC Classification Schedule Outlines. One is reminded of Cutter's third "means" to subject access in library catalogs: a "classified subject table." Two university-based experiments are exploring the use of state-of-the-art graphical interfaces for information retrieval. These design experiments involve the use of high-resolution VDUs and advanced graphics software which runs on powerful microprocessor AI workstations. They represent the furthest advances in experimental design and prototype building aimed at improving user- system dialogue interaction, orientation maintenance, and the contextual presentation of subject concepts and terminology.

The KIM (Knowledge and Information Mapping) project underway at the University of Aberdeen, Scotland, includes a trial construction and evaluation of a graphical thesaurus to be used as an interface to document retrieval systems. The work is being carried out on a Xerox 1186 workstation, and the software is written in LISP. Duncan and McAleese, the project investigators, explain: "We hope to be able to demonstrate that a graphical approach to the description of tasks or problems faced by users allows a more intelligent interaction between user and system than a dialogue or query-answer (linear) approach."[76]

The KIM researchers recognize that one of the major difficulties experienced by searchers is describing their information needs in system terms. Duncan and McAleese reject the design assumptions which underlie conventional online query systems: "Our contention is that to ask a user to describe in text form something which is at best vaguely conceived, is unreasonable, and will inevitably lead to disappointment."(76)

KIM researchers believe that a thesaurus is a valuable tool for enabling users to make their needs known to the system. However, traditional thesauri with their linear, alphabetical arrangements and one-dimensional concept relation-



Use arrow keys to scan the list. Use ENTER to go back to the match screen.

Author: Delano, Alonzo, 1806-1874.

Title: Across the plains and among the diggings, by Alonzo Delano; a reprint of the original edition with reproductions of numerous photographs taken by Louis Palenske and foreword and epilogue by Rufus Rockwell Wilson. New York, Wilson-Erickson, Inc., 1936.
xviii, 192 p. incl. front., illus. 30 cm.

Notes: First edition, 1854, has title: Life on the plains and among the diggings.

Subjects: Indians of North America--The West.
Overland journeys to the Pacific.
The West--Description and travel.

Help (F2) — Map (F3) — Expert Aids (F4) — Previous Screen (F9) — Match List (F10)

**Figure 36.** *LaserGuide Shelf Browse Screen*

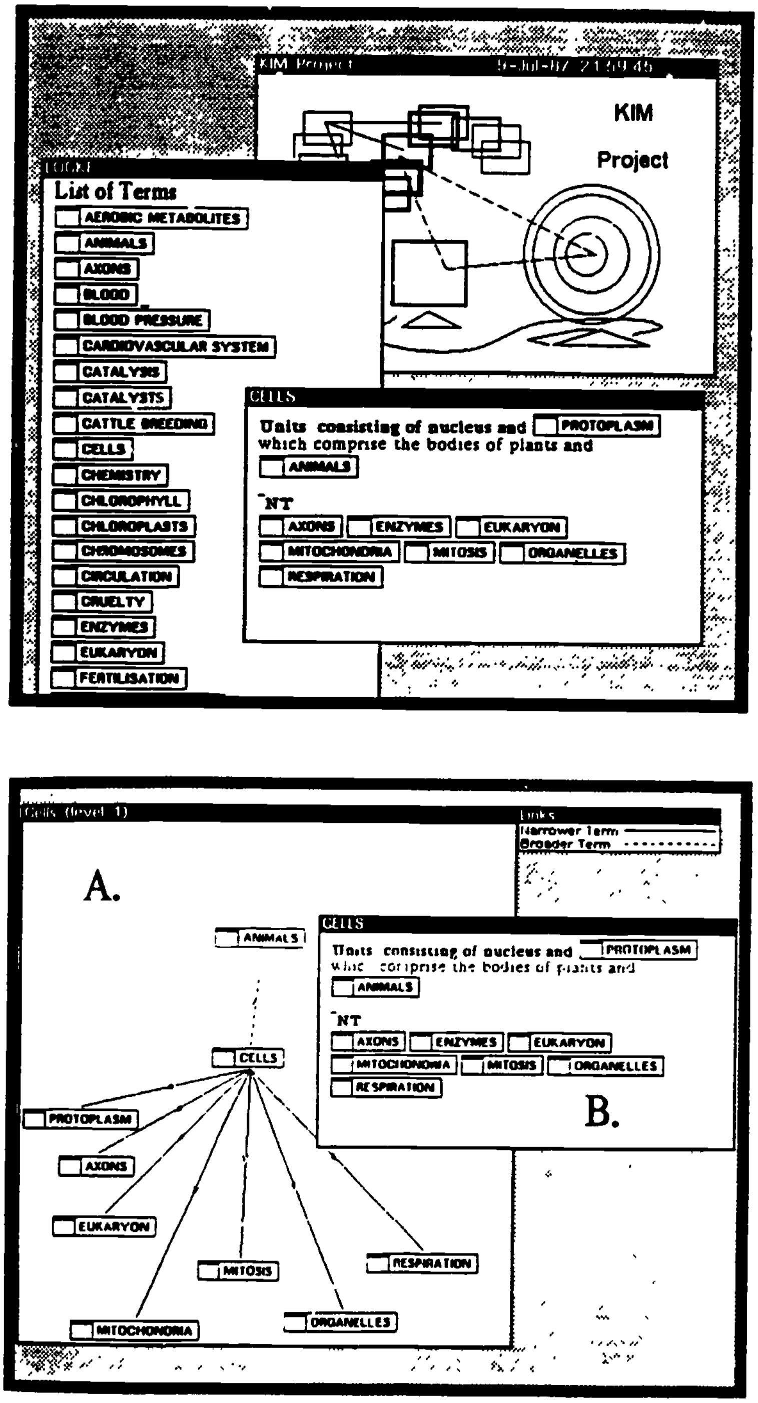
ship presentations require some training to use and may not be effective in a cognitive sense. That is, due to the complexity of thesaurus construction, it may be difficult for users to grasp a variety of term relationships just by viewing text labels like "broader" and "narrower", and the inherently ambiguous "related." Also, a given user may not have sufficient knowledge of a subject area to effectively explore a thesaurus addressed to that subject or field.

An alternative to list presentations is to display the thesaurus in graphical form as a concept map (See Figure 37). A concept map can serve as an index to the IR system by presenting related concepts as nodes in a network and graphically displaying the relationships between them. The network thus becomes the graphical equivalent of a thesaurus. This provides a visual environment of concept-term relationships that is easier to understand. The user can discover the environment of terms that were previously unclear or unknown. Early evaluations of KIM "suggest that there are learning advantages in using a graphical rather than a linear approach to the presentation of material in study and information systems."[77]

Traditional thesauri, printed or online, give the user an initial impression of being alphabetical lists of keywords or key phrases. Closer inspection and training may reveal the structural complexity of a thesaurus and the concept relationships it employs. The complex, intentional, often implicit linkages present in thesauri can be made explicit and more immediately rendered meaningful if the terms and their relationships are displayed in graphical form. This proposition is being tested in the KIM project with a variety of graphical thesaurus and concept network maps.

Duncan and McAleese note three advantages in this approach: 1) it allows the use of relationships other than the standard BT, NT, and RT; "related term" is ambiguous and may add unnecessary complexity to the user's vocabulary problem, 2) a map can be changed and different levels of detail or focus can be provided (e.g., to provide "street" or "terrain" knowledge), and 3) the visual impact of a graphical thesaurus has inherent cognitive and psychological advantages. Figure 37 (fourth screen) also shows how the searcher could view both portions of the graphic thesaurus and linked citations from the bibliographic database.

The I$^3$R prototype expert IIR system being developed under the guidance of Croft at the University of Massachusetts uses similar AI-workstation technology. Residing in a parallel-processing, distributed system architectural environment, I$^3$R supports combinable, multiple expert assistance modules designed to aid with all aspects of the information retrieval process.[78] Boolean and probabilistic retrieval options are available to the searcher. The prototype also includes an "expert" browsing facility which gives the user a means of navigating through the database from network maps of term or document relationships to document citations or full text (See Figure 38). Graphics software and windowing techniques are employed to manage the user displays.



List of Terms

- AEROBIC METABOLITES
- ANIMALS
- AXONS
- BLOOD
- BLOOD PRESSURE
- CARDIOVASCULAR SYSTEM
- CATALYSIS
- CATALYSTS
- CATTLE BREEDING
- CELLS
- CHEMISTRY
- CHLOROPHYLL
- CHLOROPLASTS
- CHROMOSOMES
- CIRCULATION
- CRUELTY
- ENZYMES
- EUKARYON
- FERTILISATION

CELLS

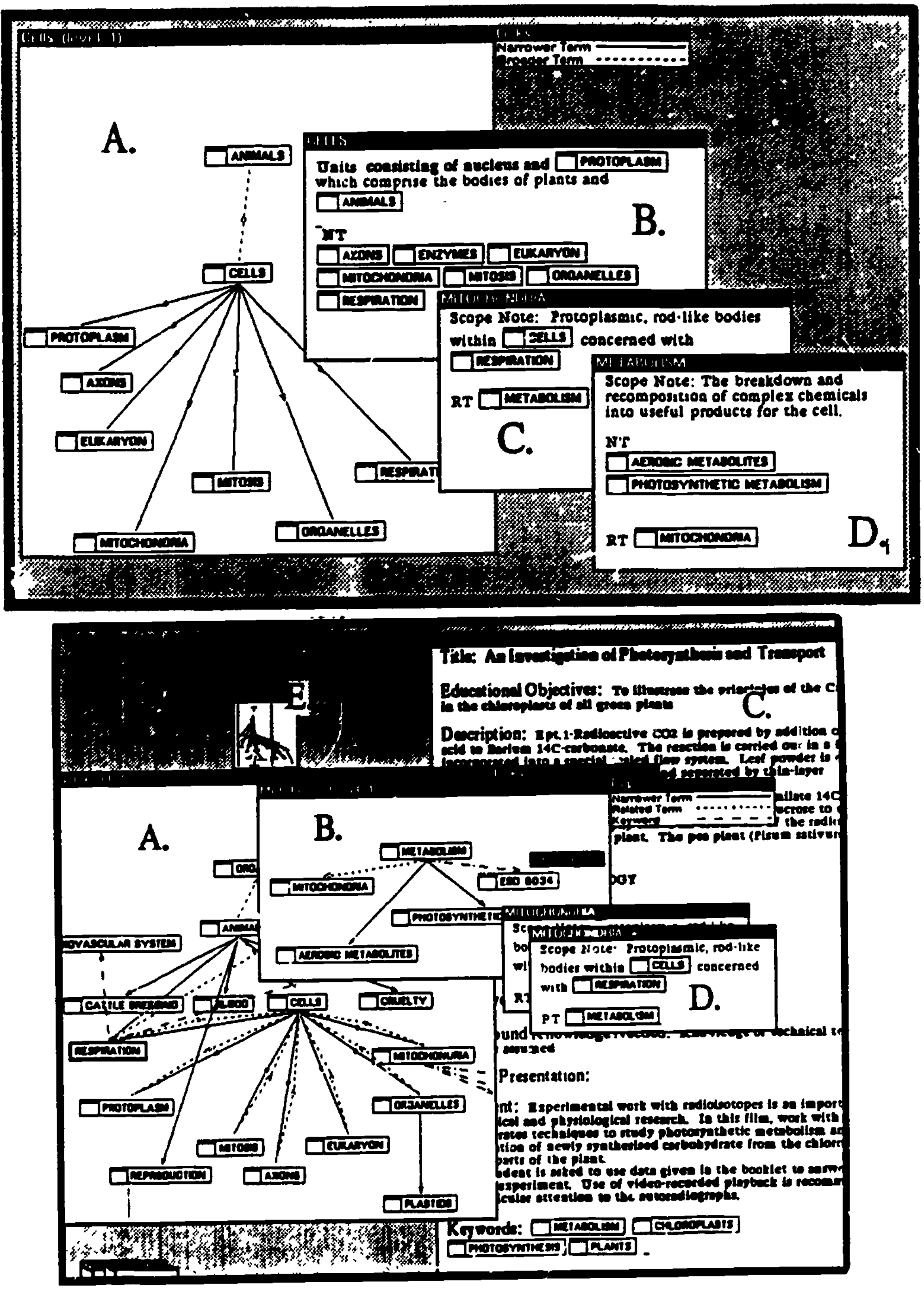
Units consisting of nucleus and PROTOPLASM which comprise the bodies of plants and ANIMALS

NT

AXONS ENZYMES EUKARYON MITOCHONDRIA MITOSIS ORGANELLES RESPIRATION

**Figure 37.** *KIM Project, University of Aberdeen*

BEST COPY AVAILABLE



Units consisting of nucleus and PROTOPLASM which comprise the bodies of plants and ANIMALS

Scope Note: Protoplasmic, rod-like bodies within CELLS concerned with RESPIRATION

RT METABOLISM

Scope Note: The breakdown and recomposition of complex chemicals into useful products for the cell.

NT AEROBIC METABOLITES PHOTOSYNTHETIC METABOLISM

RT MITOCHONDRIA

Title: An Investigation of Photosynthesis and Transport

Educational Objectives: To illustrate the principles of the C in the chloroplasts of all green plants

Description: Ept.1-Radioactive CO2 is prepared by addition o acid to Barium 14C-carbonate. The reaction is carried out in a incorporated into a special sealed flow system. Leaf powder is ed separated by thin-layer alliate 14C ucrose to the radi plant. The pea plant (Pisum sativum

Scope Note: Protoplasmic, rod-like bodies within CELLS concerned with RESPIRATION

PT METABOLISM

und knowledge needed: knowledge of technical assumed

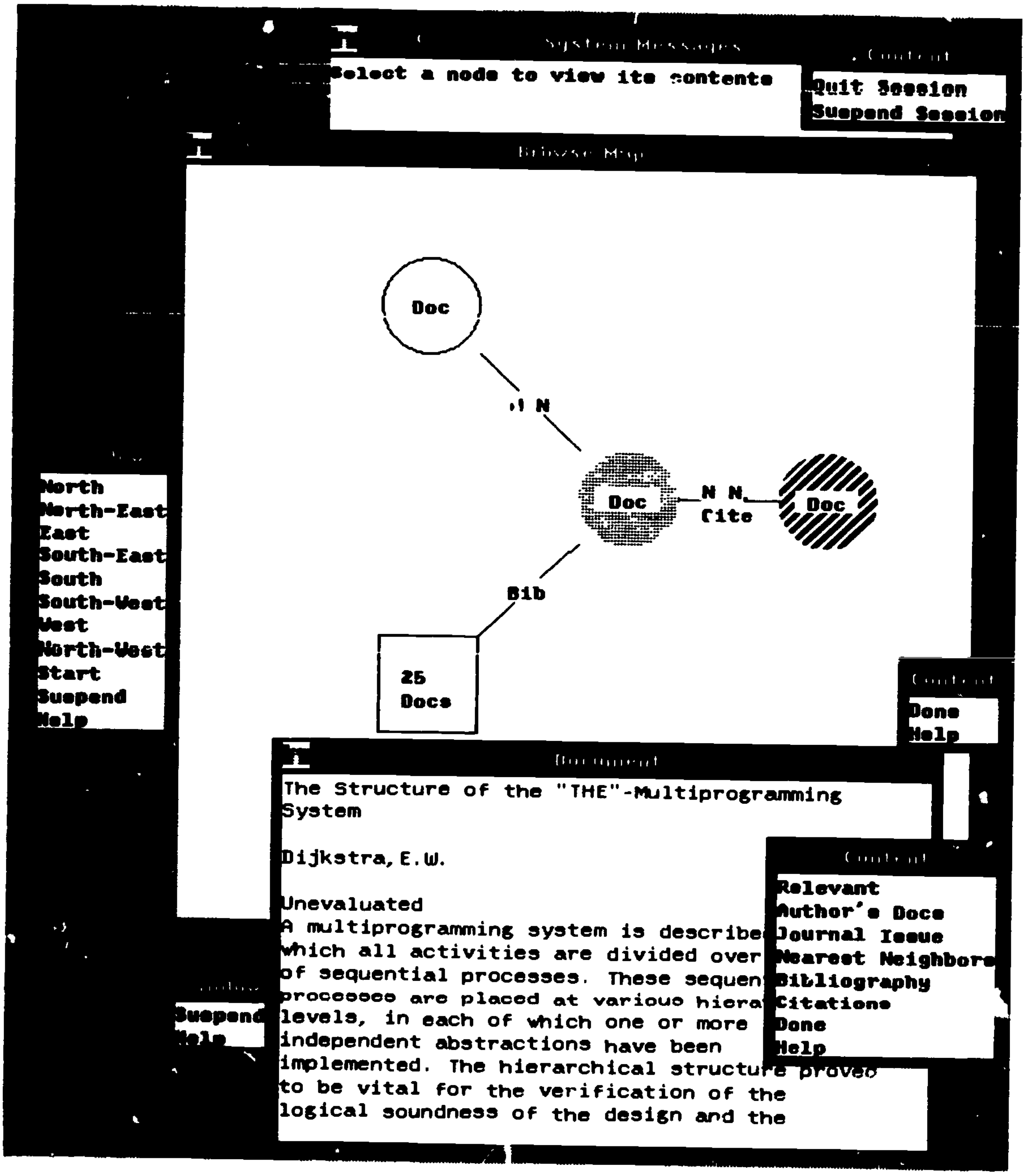
Presentation:

nt: Experimental work with radioisotopes is an import ical and physiological research. In this film, work with rates techniques to study photosynthetic metabolism a tion of newly synthesised carbohydrate from the chlor parts of the plant. udent is asked to use data given in the booklet to answ experiment. Use of video-recorded playback is recomm icular attention to the autoradiographs.

Keywords: METABOLISM CHLOROPLASTS PHOTOSYNTHESIS PLANTS

**Figure 37.** *KIM Project, University of Aberdeen (Continued)*

BEST COPY AVAILABLE



Select a node to view its contents

The Structure of the "THE"-Multiprogramming System

Dijkstra, E.W.

Unevaluated

A multiprogramming system is describe
which all activities are divided over
of sequential processes. These sequen
processes are placed at various hiera
levels, in each of which one or more
independent abstractions have been
implemented. The hierarchical structure proved
to be vital for the verification of the
logical soundness of the design and the

Figure 38. $I^3R$ Browser

$I^3R$ extends the KIM graphical thesaurus mapping approach to include the entire "knowledge base." The latter may consist of a super-thesaurus of concept category frames, and documents or data entities (e.g., authors, abstracts) which are normally stored in IR systems. The knowledge base can be displayed as a network of nodes and links. Nodes may represent entities such as citations or specific data elements they contain, and graphic links can be drawn to indicate a variety of relationships between the entities. For example, "NN" in Figure 38 represents a link to one document's "nearest neighbor" in the database. NN is calculated in this case using a statistical procedure, but it could easily be defined to mean other "works by this author" or "works in this series" or "works classed with this document." The technology permits the user to choose any network node as a starting point to follow the paths at will, and to focus in at any time on items of interest (i.e., terms or document citations or any other "entities" stored in the database, such as book reviews or journal article abstracts). $I^3R$ represents yet another approach to presenting and understanding database components or entities, such as subject headings and/or thesaural relationships, as they are used in various actual knowledge contexts.

## 5.4 Intelligent Automatic and Semi-Automatic Search Strategy Selection and Retrieval Techniques

Advanced query formulation and retrieval methods pioneered and tested in IR experiments and research studies are described in some detail in Section 4 of this report. Most of these retrieval methods and techniques constitute extensions, supplements, or alternatives to conventional Boolean, exact query match and retrieval methods. Thus, they are referred to as "post-Boolean" models, approaches, and methods, etc. Much design, development and testing work has been done to make Boolean retrieval systems easier to use, more flexible, and more effective. Smarter rule-governed or probabilistic retrieval systems have been developed to provide alternatives to Boolean searching. Some operational retrieval systems support both Boolean and post-Boolean searching methods, giving their users a choice. Third-generation OPACs are beginning to incorporate these intelligent, post-Boolean search features and retrieval methods. All the systems cited here are operational systems. Most utilize a combination of semi-automatic and automatic intelligent query matching and retrieval techniques.

### Rule-governed search strategy selection

Other features of OKAPI-86 have been described elsewhere in this report. An earlier version, OKAPI-84, was the first operational/developmental OPAC designed by the research team at the Polytechnic of Central London to be placed in a library for actual use and evaluation. The purpose of the OKAPI ongoing research, largely funded by the British Library Research & Development Department, is to test the applicability of findings from research in interactive computing, cognitive psychology, and information retrieval to OPAC design and development. OKAPI-84 was the prototype system and remains as the foundation of later versions.[79]

Two types of searching were permitted in OKAPI-84: "Books about something" or "Specific books" (See Figure 39). User's subject search words were looked up in an index containing words from titles, corporate names, and subject headings. Search strategy or "path" selection was governed by a simple set of rules in "search decision trees." The system would first attempt an implicit Boolean AND search. If no matches resulted, a weighted-term, "quorum" search was automatically performed. Search words were assigned weights based on the frequency of their occurrence in the database (uncommon words would get higher values), and the system searched for records which contained fewer than all the words in the user's query (all words minus one, all words minus two, etc.). Retrieved records were assigned a ranking value equal to the sum of the weights of the words by which they were indexed. Records with the highest rank value were listed first in the output.

Different search path rules were used for specific item searches. Alternatives were triggered according to the type of data element entered and the results of previous searches by the user. For example, if a title phrase search produced no match, OKAPI-84 would look for the titles containing the search words in any arrangement. From this point, the title search proceeded according to the programmed *subject* search rules and procedures.

Subject searching on OKAPI-84 seems to have been more successful than on conventional OPACs. A member of the research team studied a sample of subject searches captured from the transaction logs. Although only an estimate (the original searchers were not identified or consulted), it was judged that only 25% of the subject searches failed *on the first attempt.* The team points out that the *session* failure rate would have been even lower.

The LIBERTAS OPAC, developed and marketed as part of an integrated library system by SWALCAP Library Services (See Figure 12), supports sequenced-list browsing, and full Boolean retrieval via commands for searchers who wish to bypass the menu-structure. The menu-selected "Subject keyword" search accepts query input in natural language. LIBERTAS employs a search decision tree, combinatorial search facility much like OKAPI's. The aim of the system is to "rescue" initial subject searches that fail to produce any matches. Words from both title and subject fields are included in the subject search index. If no record contains all the search words, the system "looks for items described by as many as possible of them." (From a LIBERTAS help display)

It was previously noted that BiblioFile's CD-ROM Intelligent Catalog uses an automatic "try harder" search rule when no exact match for a user-entered phrase can be found. If initial syntactical restructuring of the phrase (e.g., flipping inverted terms) does not result in a match, the system then performs the search as a Boolean AND query. LaserGuide, the GRC CD-ROM OPAC (See Figure

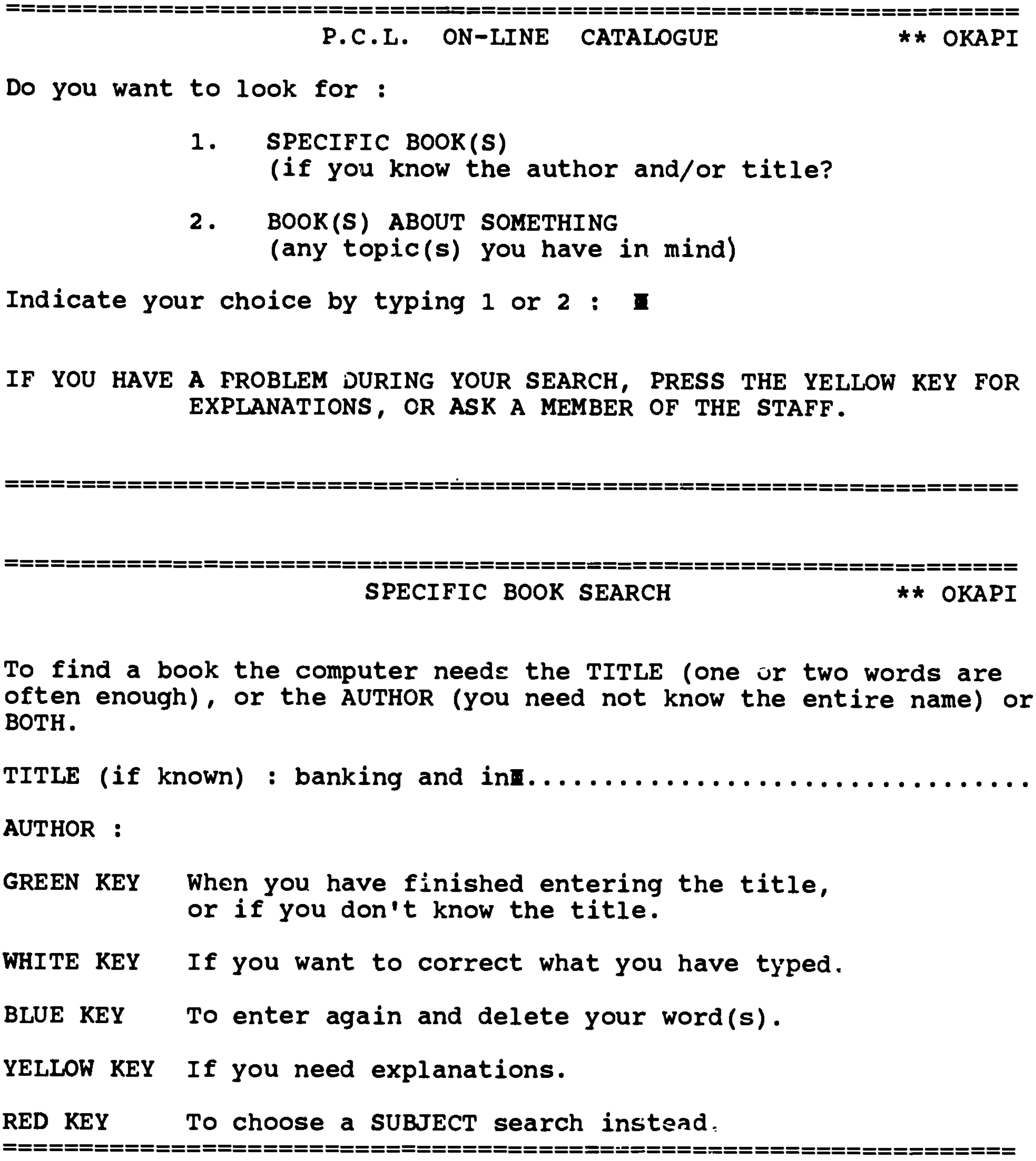
26), offers a flexible subject searching facility and employs a one or two-step procedure as necessary to find records that match the user's query. Initial subject queries are performed as keyword, Boolean AND searches on all of the words of all subject headings in the database. If a subject search results in no matches, LaserGuide automatically expands the search to all words in the stored bibliographic records. If no matching words are found, the system displays words that begin with the same characters.

### Weighted-Term Probabilistic Retrieval With Relevance Feedback and Output Ranking

CITE, SIRE, OCLC's CD-ROM reference retrieval product, and OKAPI-86 are OPAC/IR systems previously noted in this report that employ some combination of these post-Boolean methods. OKAPI-86 (See Figure 24) permits only subject searching by its library users. All searches are performed by the system on a best match, combinatorial basis. A match on fewer than all of the query words may retrieve items, but records which have all the query words are displayed first and described as "matching your search closely." Search terms are weighted, so "methods" would get a low "importance" weight compared to the weight assigned to a rare word like "yodelling". In the search for "methods of yodelling," records containing both "methods" and "yodelling" are output first, followed by records indexed under "yodelling" alone. Records indexed under "methods" but not "yodelling" probably would not be retrieved, as postings thresholds are applied to "stoplist" a highly posted word. Additional rules define what weight a record has to achieve in the combined score of the weights of its matched terms for it to be retrieved as a relevant match. OKAPI's designers set these levels on the basis of experience and log analysis to achieve a balance of recall and precision in most search results. The assumption is, for example, that few books are indexed under "yodelling" and the searcher may wish to learn about all of them. Evaluation results to date of OKAPI's subject searching performance were reviewed earlier in this section of the report.

SPRILIB is a new, state-of-the-art third-generation OPAC now in operation at the Scott Polar Research Institute Library at Cambridge University (UK).[80] SPRILIB provides full probabilistic retrieval with query expansion, relevance feedback, and ranked output, as well as conventional Boolean retrieval. Probabilistic retrieval is applied to queries entered in free text, natural language. The OPAC may be used in either menu mode or full command mode (23 commands). Function keys are used to invoke menu selections. Boolean retrieval may be combined with probabilistic retrieval, making it possible to weight terms in a Boolean results set and output individual citations in order of decreasing relevance. (*Note*: The experimental SIRE system was one of the first to apply term weighting and document ranking to Boolean retrieval.[36])

SPRILIB uses a software package called Muscat developed at Cambridge University for use in museum cataloging database management, and information



```
=================================================================
                   P.C.L.  ON-LINE  CATALOGUE          ** OKAPI

Do you want to look for :

            1.    SPECIFIC BOOK(S)
                  (if you know the author and/or title?

            2.    BOOK(S) ABOUT SOMETHING
                  (any topic(s) you have in mind)

Indicate your choice by typing 1 or 2 :  █


IF YOU HAVE A PROBLEM DURING YOUR SEARCH, PRESS THE YELLOW KEY FOR
            EXPLANATIONS, OR ASK A MEMBER OF THE STAFF.


=================================================================
```

```
=================================================================
                      SPECIFIC BOOK SEARCH             ** OKAPI


To find a book the computer needs the TITLE (one or two words are
often enough), or the AUTHOR (you need not know the entire name) or
BOTH.

TITLE (if known) : banking and in█................................

AUTHOR :

GREEN KEY    When you have finished entering the title,
             or if you don't know the title.

WHITE KEY    If you want to correct what you have typed.

BLUE KEY     To enter again and delete your word(s).

YELLOW KEY   If you need explanations.

RED KEY      To choose a SUBJECT search instead.
=================================================================
```

**Figure 39.** *OKAPI-84*

retrieval. The system runs on a University IBM 3084Q2 mainframe computer. The bibliographic database consists of standard, full MARC (UK) catalog records for materials in the library's collection. The collection is classified according to UDC and the codes are indexed, searchable, and may be used for automatic query expansion. Inverted index files are used for searching and matching. Muscat supports a number of types of file organization (e.g., ordinary text files, indexed sequential files of bibliographic records). The Muscat database is maintained as a B-tree. Individual term postings (word occurrence) data is used in the weighting and ranking algorithms.

Using either dialogue mode, searchers may conduct an author search or a subject/title search (the latter two are consolidated). Keywords extracted from document titles are indexed for the subject/title search. Likewise, words are selected for matching by the system from the user's natural language query. A word stemming routine is applied in both cases. After a query is entered by the user, SPRILIB performs a combinatoric "best match" search and outputs retrieved citations in order of decreasing probability of relevance. A query statement that exactly matches a title phrase will result in that title being displayed first. Each displayed citation ends with the question, "Relevant?"; the user may answer YES or NO using function keys. These judgments are subsequently used for query expansion and iterative searching if this is desired by the user.

The SPRILIB command mode permits the user to search either by the natural language, probabilistic method or the Boolean method.

For example,

query penguin breeding habits

invokes a probabilistic query, while,

boolean 'penguin' and 'breeding' and 'habits'

invokes a Boolean query. Terms may be added to a probabilistic query at any stage. There are three ways to do this: by selecting a term from an index browsing list, adding a term with the "add" command, and by using the "terms" command which is a relevance feedback method of semi-automatically expanding the query in progress. Terms may also be deleted. If the user's query contains both Boolean and probabilistic components, the results set includes only documents which satisfy the Boolean query, but these are output ranked as in a probabilistic search.

The term weighting and document ranking algorithms are described in the publication cited. Generally speaking, the weighting of a term is based on two factors: the frequency of the term in the document collection as a whole, and its frequency in the set of relevant retrieved documents (those marked "YES" by the user). The latter measure is used in the query expansion and redefined search.

Research has shown relevance feedback can improve subject retrieval. As in CITE, terms or other subject indicators are automatically selected from retrieved documents judged by the user to be relevant. These terms can then be used to automatically (no user review) or semi-automatically expand the search. SPRILIB permits the user to view the candidate terms, including UDC codes, the system has selected (and often stems) for query expansion in the continuation subject search process. The terms are displayed in ranked order. For each candidate search term the user is asked to respond YES or NO. Any approved terms are then used in the redefined, continued, subject search.

At the time of this report, formal evaluations of SPRILIB are not complete, as the OPAC has been in place only about one year. The SPRI Librarian reports, however, that there is evidence of retrieval improvement after a second feedback cycle, depending, of course, on which new terms the searcher approves for query expansion. Users seem to be very satisfied with the alternative to Boolean searching: "It is easy to convince users of the utility of relevance feedback when one can present relevant documents which are not indexed by any of the terms in the original query formulation."[81]

An intelligent retrieval system very similar to SPRILIB has been developed as a demonstration system at Canberra, Australia.[82] As its name implies, "STATUS with IQ" consists of software added-on to STATUS, the widely used document and text retrieval commercial software product. STATUS with IQ can run on a variety of computers, including IBM MVS/TSO or VM/CMS machines, and DEC VAX machines.

Status with IQ supports both Boolean and weighted-term searching with relevance feedback (it employs a somewhat different weighting and ranking algorithm). Setting it apart from SPRILIB, the Australian system employs an automatic "natural language analyzer" to interpret user queries. This NLP module not only stems query words but also performs syntax-based decomposition and restructuring of any phrases entered in the natural language query. One set of rules governs the extraction and editing of words and phrases from queries. Another set of rules directs work on these "categories" to generate the search terms. For example, capitalized phrases ("United States") are generated "as is" to become search terms. More complex phrase term generation is performed using a syntax pattern matching facility. The system does permit the user to edit the reconstructed query prior to performing the search. Pape and Jones note that automatic generation of search terms based only on syntactical analysis of users' query terms can produce some odd results.[83] They do not recommend using a syntactical language processor as a front end to Boolean queries unless it is used in conjunction with a controlled vocabulary that has been used to index the documents in the retrieval system.

## 6. Conclusion

This report has reviewed the major subject searching problems encountered in the use of today's OPACs and end-user information retrieval systems, research and development issues and advances, and intelligent system design challenges for improving subject access to bibliographic databases. Some problem and issue areas have been discussed at greater length than others (e.g., vocabulary matching/translation) because they are more central to the challenge and tasks we face in improving subject access and retrieval capabilities in the next generation of operational, widely-available OPACs.

*Insights* acquired about critical OPAC use and design problems, design *principles* based on research and experience, and design *recommendations* for improving OPACs and IR systems all boil down to pretty much the same thing: they are different ways of answering the question, What should be done now in system redesign or replacement efforts to make our subject retrieval systems more usable and more effective, and, thereby, to greatly improve the lot of today's subject searchers?

This study suggests nothing less than a fundamentally revised perspective on users' subject searching needs and OPAC design to respond to those needs. The fact that users bring different types of subject search needs and different subject search requirements to OPACs cannot be ignored any longer. A single search interface and a single retrieval model (e.g., Boolean) are not adequate for the variety of searchers and search needs that exist. "Moving up" to conventional Boolean retrieval enhanced with a friendly interface has not bought OPAC users much retrieval effectiveness.

There is a large category of search needs and behavior for which the Boolean exact match query model is simply wrong. Effective search and retrieval methods must also be provided for the subject browser, and for searchers who cannot, or do not wish to, precisely describe their information need or problem in advance of the activation of the retrieval process. Librarians have long understood the differences between "known-item" searches and "subject" searches in the catalog. Now, we need to equally appreciate different kinds of subject searches and searching requirements. Alternative design models and retrieval methods which support these separate needs and modes of searching behavior should be integrated into improved subject OPACs. Users should be given a choice of 1) how they want to access the subject catalog, 2) how they wish to use it (to match searching styles or requirements), and 3) how much automatic and/or semi-automatic assistance they want from the system. "Known-subject" searchers, that is, searchers who can articulate their queries precisely in the language of the system, may be served best by a fully-automatic, "rifle-shot" Boolean retrieval system enhanced with ranked-document output capabilities. The majority of subject searchers probably need a more interactive subject searching approach which gives opportunity for term and document appraisal and relevance feedback *during* the search process. Determined explorers and the just plain curious need a flexible, rich, contextual subject search and browsing mode which offers plenty of navigation and trail blazing options.

## CONCLUSION

### Applicability of Intelligent Retrieval Methods to a Future LC Information Retrieval System

To suggest OPAC/IR system design and development priorities and/or methods of implementation is beyond the scope of this report. However, a better understanding of the critical problems users face will surely move some feasible subject access "solutions" to the top of our priority lists. Matching and retrieval techniques pioneered in IR and AI research that are known to improve retrieval performance should be incorporated in future OPACs. To summarize, these include natural language query processing; direct or indirect mapping/linking of free text terms to terms in the controlled vocabulary used to index documents; flexible, heuristic retrieval strategies; but, primarily, probabilistic retrieval with weighted-term, combinatorial searching and the ranking of output; and "user-engaged" relevance feedback procedures for automatic query expansion and modified search strategies.

Assessing the applicability of various advanced intelligent retrieval and interface methods to a future Library of Congress information retrieval system depends very largely on what assumptions are made about the data constituents, size, architecture, and projected users of that system. Does the "future" mean two years from now, or five-to-ten years? Since I possess little detailed information on these matters my assumptions are few and can be stated simply.

A future LC information retrieval system

- will include very large bibliographic (MARC and non-MARC) files,
- will employ multiple thesauri (lacking a common structure and vocabulary) to govern entries in some of the bibliographic files,
- will incorporate a traditional database structure which utilizes inverted-file indexes,
- will provide a Boolean query approach as one of its search modes,
- will utilize multiple, powerful processors for carrying out different storage, update, and retrieval tasks, perhaps in a distributed processing configuration,
- will use intelligent, microcomputer-based workstations for searchers' "terminals", and
- will provide bibliographic access service to a variety of users, including untrained, infrequent searchers, both scholar/researchers and members of the general public.

# CONCLUSION

From an idealistic point of view, all of the design developments reviewed in the preceding section represent *desirable* directions for LC to pursue in planning, designing, and building its future IR system. In this sense, they are all applicable in some degree to LC's efforts to build an advanced, state-of- the-art intelligent IR system. From a practical point of view, at least for the short term, desirability must compete with *feasibility*. However, I do assume a common desire on the part of LC's system planners to take a giant step beyond the present systems, SCORPIO and MUMS, to improve IR system usability and retrieval effectiveness.

Methods and algorithms proven to be efficient and effective exist to improve present-day OPAC/IR system interaction and retrieval performance. They address four "requirements" areas of a future LC retrieval system: 1) easier, system-assisted search entry and query formulation; 2) automatic assistance with failed searches through query expansion/reformulation or changed search strategy; 3) improved, extended-Boolean retrieval methods; and 4) data linking and windowing techniques to bring thesaural terms and relationships into the context of the actual search process.

The state-of-the-art review included examples of Boolean systems which accept user queries in natural language. These systems do not require the user to learn a complex command syntax and to become expert in the use of Boolean logic operators. Their capabilities range from simple parsing of user search input against word stoplists or cross-reference term dictionaries and the automatic insertion of logical operators between the accepted words, to "intelligent" word stemming, automatic spelling corrections, and simple syntactical manipulations of composite terms (e.g., word inversion, noun-adjectival form transformations). These techniques relax the search entry requirements placed on the user and increase the probability of finding word/term matches in the system's indexes. Much of the software for these routines could reside in the microcomputer-based intelligent workstation.

Software residing in the searchers' terminals at the University of Illinois at Urbana-Champaign handles the reformulation of search strategy when a subject heading phrase-match search fails. The search is switched to a title-keyword search targeted to an entirely different database. This change is explained to the user, who is then guided to find a relevant subject heading with which to continue the search for relevant items. Several OPACs and IR systems, such as OKAPI, LIBERTAS, and Bibliofile's, automatically implement alternative search strategies when the first attempt fails. Many systems at least suggest to the searcher alternative strategies for improving retrieval results. For a new, untrained searcher, the system could operate in automatic "expert" search formulation/strategy selection mode. An experienced searcher could be presented with the choice to accept or not accept the automatic mode.

Extended-Boolean retrieval methods can be used quite effectively with conventional inverted file structures to enhance the retrieval performance of

Boolean search systems. Efficient term weighting algorithms exist which exploit data already in the inverted files (e.g., word occurrence frequency data) to support both "partial" and "best match" retrieval and the ranking of retrieved documents. Short of replacing Boolean systems entirely (as OKAPI does), or providing both Boolean and probabilistic, relevance-feedback search approaches in a single system (as SPRILIB does), weighted-term retrieval can be "added-on" to significantly improve the retrieval performance of inverted file, Boolean systems.

The simplest alternative to Boolean query-matching is "quorum" searching. Matches are found for as many of the user's search words as possible. A 4-word search processed initially with Boolean ANDs does not fail if a record can be found containing three of the words, and so on. To supplement quorum searching, a simple term "weighting" scheme, on which retrieved citations can be rank-ordered, judges the first word or words entered by the searcher as having more "value" or importance for the search aim. Relevance feedback mechanisms, through which judgments of relevant terms or documents may be supplied by the user during the search, can be incorporated in conventional Boolean retrieval systems to refine/limit previous search results, or to serve as the basis for new or extended searches conducted by the system.

Restructuring and/or integrating LC's several thesauri is, no doubt, a massive undertaking. However, proper database design and data management software could support the establishment of linkages and "trails" both among terms in these thesauri and entries in the bibliographic files. This would permit navigation by the searcher among terms in differen*t* thesauri, as well as bibliographic record-to-bibliographic record navigation. Some "vocabulary- switching" algorithms have been shown to be effective in support of this capability, but it is not necessary to implement these algorithms to create term-browsing trails using shared properties of terms found in two different thesauri, for example, associated words or phrases found in titles of works both vocabularies have been used to index. The searcher could be guided not only to text terms associated in that database with terms from a controlled vocabulary (or vice versa), but also to terms in a related thesaurus (e.g., MeSH, AAT, etc.). The terms in the related thesaurus may be more timely and discipline-specific than LCSH, and thus may provide the searcher better search vocabulary for identifying additional relevant materials.

These search trail-based linkages betwe*e*n different vocabularies could be established in the database by the database management software using data already existing in the stored bibliographic records (e.g., titles, names, call numbers). Display windowing techniques, now commonplace in microcomputer software products, could be used to reduce vocabulary confusion and to keep the searcher oriented while moving from vocabulary to vocabulary, vocabulary to citation, citation to citation, and so on.

Some AI-based expert information retrieval systems have recently appeared (e.g., the TOME Searcher), but their application to OPAC retrieval may have built-in obstacles too large to overcome. Thus far, these knowledge-based, rule-governed, expert IR systems have been effective and feasible only for specialized subject databases and a narrow range of subject information needs. They require a rich knowledge base consisting of subject reference knowledge, as well as knowledge about how it may be used by experts to solve an immense variety of problems in a particular field. This challenge requires a very large amount of domain-specific knowledge and rules for its problem-solving use even in a small collection of documents. Library catalogs, on the other hand, cover many subject fields and support an infinite variety of information needs. In this environment, a more feasible "expert" system approach is one that helps- out primarily "up front" in the selection of the best search vocabulary for retrieving potentially relevant documents (e.g., ERLI's NLQP and Mischo's University of Illinois experiment). Once into an OPAC at hand, an OPAC that is hospitable and richly suggestive of search and retrieval options, the direct end user is probably expert enough to decide what he wants and when he has found it.

In the design of new information retrieval systems, at a time when computer technologies support parallel processing, distributed "expert" architectures, and intelligent workstations/terminals, a fresh perspective on subject retrieval is needed. Intelligence must reside throughout the system, from the request/presentation interface to database modelling, structure, and content. More "windows" to the system and the database can be opened up at the VDU; improved matching and retrieval methods have been tested and are available for adoption; and databases can be modelled and structured to support enhanced retrieval, browsing and navigation, rather than cataloging activities. We are faced with a pleasant irony: many additional "automatic" techniques can be incorporated to make IR systems more effective and, at the same time, enable them to become more cooperative and engaging.

## References

1. Bates, Marcia J. "Subject Access in Online Catalogs: A Design Model." Journal of the American Society for Information Science, 37:6(November 1986); 357-376.

2. Mischo, William H. and Lee, Jounghyoun. "End-User Searching of Bibliographic Databases." In: Annual Review of Information Science and Technology (ARIST), Vol. 22, 1987; 227-263.

3. Mischo, William H. and Lee, Jounghyoun. *Op. Cit.*

4. Cutter, Charles Ammi. Rules for a Dictionary Catalog, 4th ed., Washington, D.C.: Government Printing Office. 1904; 12.

5. Bates, Marcia J. *Op. Cit.*

6. Gerrie, Brenda. Online Information Systems. Information Resources Press, 1983. (Distributed by Gothard House Publications: Henley-on-Thames, Oxon. RG9 1AJ, UK.)

7. Pritchard, Sarah. "Subject Access in the LOCIS Environment." Automation at the Library of Congress: Inside Views, edited by Suzanne E. Thorin. Washington, D.C.: Library of Congress Professional Association. 1986; 27-30.

8. Markey, Karen. Subject Searching in Library Catalogs: Before and After the Introduction of Online Catalogs. (OCLC Library, Information, and Computer Science Series) Dublin, Ohio: OCLC Online Computer Library Center. 1984.

9. Markey, Karen. *Op. Cit.*

10. Markey, Karen. Dewey Decimal Classification Online Project: Evaluation of a Library Schedule and Index Integrated into the Subject Searching Capabilities of an Online Catalog: Final Report to the Council on Library Resources. Dublin, Ohio: OCLC, Report no. OCLC/OPR/RR-86/1, 1986.

11. Cutter, Charles Ammi. *Op. Cit.*

12. Bates, Marcia J. *Op. Cit.*

13. Markey, Karen. "Users and the Online Catalog: Subject Access Problems." In: The Impact of Online Catalogs, edited by Joseph R. Matthews. New York: Neal-Schuman Publishers, Inc. 1986; 35-69. Karen Markey, Subject Searching in Library Catalogs: Before and After the Introduction of Online Catalogs. Dublin, Ohio: OCLC, 1984.

14. Knipe, Nancy. "Hands-on: User-directed System Evaluation." Conference on Integrated Online Library Systems, 23-24 September 1986, St. Louis; Proceedings. David C. Genaway, ed. Canfield, Ohio: Genaway & Associates, 1987; 327-40.

15. Markey, Karen. *Op. cit.*

16. Kranich, Nancy C., *et al.* "Evaluating the Online Catalog from a Public Services Perspective: A Case Study at the New York University Libraries." In: The Impact of Online Catalogs, edited by Joseph R. Matthews. New York: Neal-Schuman Publishers, Inc. 1986; 89-140.

17. Chan, Lois Mai. "Library of Congress Classification as an Online Retrieval Tool: Potentials and Limitations." Information Technology and Libraries, 5:3(September 1986): 181-192.

18. *Ibid.*, p. 188. See also: Croft, W. B. "Incorporating Different Search Models Into One Document Retrieval System." ACM SIGIR Forum, 16(1981); 40-45.

19. Markey, Karen. Dewey Decimal Classification Online Project: Evaluation of a Library Schedule and Index Integrated into the Subject Searching Capabilities of an Online Catalog: Final Report to the Council on Library Resources. Dublin, Ohio: OCLC, Report no. OCLC/OPR/RR-86/1, 1986.

20. Markey, Karen. Subject Searching in Library Catalogs: Before and After the Introduction of Online Catalogs. (OCLC Library, Information, and Computer Science Series) Dublin, Ohio: OCLC Online Computer Library Center. 1984.

21. Pritchard, Sarah. *Op. Cit.*

22. Bates, Marcia J. *Op. Cit.*

23. Mandel, Carol A. and Herschman, Judith. "Online Subject Access - Enhancing the Library Catalog." Journal of Academic Librarianship, 9:3(July 1983); 148-155.

24. Mandel, Carol A. and Herschman, Judith. *Op. Cit.*

25. Bates, Marcia J. *Op. Cit.*

26. Cochrane, Pauline A. (Personal conversation with the author)

27. Bates, Marcia J. *Op. Cit.*

## REFERENCES

28. Doszkocs, Tamas E. "From Research to Application: The CITE Natural Language Information retrieval System." Research and Development in Information Retrieval, Gerard Salton and Hans-Jochen Schneider, eds. (Lecture Notes in Computer Science Series, 146) Berlin: Springer-Verlag, 1983.

29. Porter, Martin and Galpin, Valerie. "Relevance Feedback in a Public Access Catalogue for a Research Library: Muscat at the Scott Polar Research Institute." Program, 22:1(January 1988); 1-20.

30. Porter, Martin and Galpin, Valerie. *Op. Cit.*

31. Mischo, William H. and Lee, Jounghyoun. *Op. Cit.*

32. Mischo, William H. and Lee, Jounghyoun. *Op. Cit.*

33. Friend, Linda. "Independence at the Terminal: Training Student End Users to do Online Literature Searching." Journal of Academic Librarianship, 11:3(July 1985); 136-141.

34. Pape, D. L. and Jones, R. L. "STATUS With IQ - Escaping From the Boolean Straightjacket." Program, 22:1(January 1988); 32-43.

35. Salton, G., Buckley, C., and Fox, E. A. "Automatic Boolean Formulations in Information Retrieval." Journal of the American Society for Information Science, 34:4(July 1983); 262-280.

36. Noreault, Terry, Koil, Matthew, and McGill, Michael. "Automatic Ranked Output from Boolean Searches in SIRE." Journal of the American Society for Information Science, (November 1977); 333-339.

37. Salton, Gerard. "The Use of Extended Boolean Logic in Information Retrieval." Proceedings of the Annual Meeting of the ACM SIGMOD, Boston, MA, 18- 21 June 1984. SIGMOD Record, Vol. 14, No. 2; 277-285.

38. Belkin, N. J. and Vickery, A. Interaction in Information Systems. The British Library, 1985. (Library and Information Research Report 35.); Davies, R., ed. Intelligent Information Systems. Ellis Horwood, 1986; Gerric, Brenda. Online Information Systems. Information Resources Press, 1983. (Distributed by Gothard House Publications: Henley-on-Thames, Oxon. RG9 1AJ, UK.) ; Salton, G. and McGill, M. J. Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983.

39. Eastman, Caroline M. "An Approach to the Evaluation of Catalog Selection Systems." Information Processing & Management, 24:1(1988); 23-30.

40. Bookstein, Abraham. "Information Retrieval: A Sequential Learning Process." Journal of the American Society for Information Science, 34:5(September 1983); 331-342.

41. Bates, Marcia J. *Op. Cit.*

42. Doszkocs, Tamas E. "Natural Language Processing in Information Retrieval." Journal of the American Society for Information Science, 37:4(July 1986); 191-196.

43. Bookstein, Abraham. *Op. Cit.* See also: Bookstein, Abraham. "Probability and Fuzzy-Set Applications to Information Retrieval." Annual Review of Information Science and Technology, Vol. 20, 1985; 117-151.

44. Croft, W. Bruce. "Approaches to Intelligent Information Retrieval." Information Processing & Management, 23:4(1987); 249-254.

45. Croft, W. Bruce and Thompson, R. H. "$I^3R$: A New Approach to the Design of Document Retrieval Systems." Journal of the American Society for Information Science, 38:6(November 1987); 389-404.

46. Croft, W. Bruce and Thompson, R. H. *Op. Cit.*

47. Robertson, S. E. and Sparck Jones, K. "Relevance Weighting of Search Terms." Journal of the American Society for Information Science, 27:3 (1976); 129-146. Also: Harper, D. J. Relevance Feedback in Document Retrieval, PhD Thesis, Cambridge University, 1980; Oddy, R. N. "Information Retrieval Through Man-Machine Dialogue." The Journal of Documentation, 33:1(March 1977); 1-14; Hendry, Ian G., *et al.* "INSTRUCT: a Teaching Package for Experimental Methods in Information Retrieval. Part 1: The User's View". Program, 20:3(July 1986); 245-263.

48. Croft, W. Bruce and Thompson, R. H. *Op. Cit.*

49. Croft, W. Bruce and Thompson, R. H. *Op. Cit.*

50. Doszkocs, Tamas E. "Natural Language Processing in Information Retrieval." *Op. Cit.* See also the special issue: "The Potential for Improvements in Commercial Document Retrieval Systems." Information Processing & Management, 24:3(1988).

51. Smith, Linda C. "Machine Intelligence vs. Machine-Aided Intelligence in Information Retrieval: A Historical Perspective." Research and Development in Information Retrieval, Gerard Salton and Hans-Jochen Schneider, eds. (Lecture Notes in Computer Science Series, 146) Berlin:Springer-Verlag,1983;263-74.

## REFERENCES

52. Williams, Martha E. "Highlights of the Online Database Field, Gateways, Front Ends and Intermediary Systems." In: Proceedings of the 6th National Online Meeting, New York, April 30-May 2, 1985. Medford: Learned Information, 1985; 1-4.

53. Deschatelets, Gilles. "The Three Languages Theory in Information Retrieval." International Classification, 13:3(1986); 126-132.

54. Doszkocs, Tamas E. "Natural Language Processing in Information Retrieval." *Op. Cit.*

55. Bates, Marcia J. *Op. Cit.*

56. Markey, Karen. Subject Searching in Library Catalogs: Before and After the Introduction of Online Catalogs. *Op. Cit.*

57. Svenonius, Elaine. "Unanswered Questions in the Design of Controlled Vocabularies." Journal of the American Society for Information Science, 37:5(September 1986); 331-340.

58. Tague, Jean. "Negotiation at the OPAC Interface." (Draft paper, to be published in 1988 monograph on OPAC Research Issues, C. Hildreth, ed., by the British Library Association Press.)

59. Bates, Marcia J. *Op. Cit.*

60. Congreve, Juliet. "Problems of Subject Access: (i) Automatic Generation of Printed Indexes and Online Thesaural Control." Program, 20:2(April 1986); 204-210.

61. Mandel and Herschman, *Op. Cit.*

62. Markey, Karen. Dewey Decimal Classification Project. *Op. Cit.*

63. Congreve, Juliet. *Op. Cit.*

64. Mandel and Herschman, *Op. Cit.*, and Tague, Jean. *Op. Cit.*

65. Tague, Jean M. "User-Responsive Subject Control in Bibliographic Retrieval Systems." Information Processing & Management, 17(1981); 149-159.

66. Walker, Stephen and Jones, Richard M. Improving Subject Retrieval in Online Catalogues: 1. Stemming, Automatic Spelling Correction and Cross-Reference Tables. London: British Library, 1987. (British Library Research Paper, 24).

67. Porter, M. F. "An Algorithm for Suffix Stripping." Program, 14:3(July 1980); 130-137.

68. Doszkocs, Tamas and Ulmschneider, John E. "A Practical Stemming Algorithm for Online Search Assistance." In: Proceedings, National Online Meeting - 1983. Compiled by Martha E. Williams and Thomas H. Hogan. Medford: Learned Information. 1983; 93-106.

69. Doszkocs, Tamas E. "From Research to Application: The CITE Natural Language Information Retrieval System." Research and Development in Information Retrieval, Gerard Salton and Hans-Jochen Schneider, eds. (Lecture Notes in Computer Science Series, 146) Berlin: Springer-Verlag, 1983.

70. Le Loarer, P. "Natural Language Approach and Online Information Retrieval Systems." Paper presented at IATUL (International Association of Technological University Libraries) Conference, July 7-11, 1986. Compiegn, France. Paris: Societe ERLI.

71. Tague, Jean. "Negotiation at the OPAC Interface." *Op. Cit.*

72. Shoval, Peretz. "Principles, Procedures and Rules in an Expert System for Information Retrieval." Information Processing & Management, 21:6(1985); 475-487.

73. Horowitz, Gary L. and Bleich, Howard L. "PaperChase: A Computer Program to Search the Medical Literature." New England Journal of Medicine, 305:16(1981); 924-930.

74. Wade, Stephen J. and Willett, Peter. "INSTRUCT: A Teaching Package for Experimental Methods in Information Retrieval. Part III. Browsing, Clustering and Query Expansion." Program, 22:1(January 1988); 44-61

75. *Op. Cit.*

76. Duncan, E. B. and McAleese R. "Intelligent Access to Databases Using a Thesaurus in Graphical Form." In: Proceedings of the 11th International Online Information Meeting, December 10-13, 1987, London; 377-387.

77. *Op. Cit.*

78. Croft, W. Bruce and Thompson, R. H. "$I^3R$: A New Approach to the Design of Document Retrieval Systems." Journal of the American Society for Information Science, 38:6(November 1987); 389-404.

## REFERENCES

79. Mitev, Nathalie Nadia, *et al.* Designing an Online Public Access Catalogue: OKAPI, a Catalogue on a Local Area Network, The British Library (Library and Information Science Report No. 39). British Library Publications Sales Unit, Boston Spa, Wetherby, West Yorkshire LS23 7BQ. Also: Mitev, Nathalie Nadia and Walker, Stephen. "Information Retrieval Aids in an Online Public Access Catalog: Automatic Search Sequencing." In Informatics 8: Advanced Computational Techniques for Information Retrieval. Proceedings of a conference organized by the Aslib Informatics Group and the BCS Information Retrieval Specialist Group, 16-18 April 1985, Wadham College, Oxford.

80. Porter, Martin and Galpin, Valerie. "Relevance Feedback in a Public Access Catalogue for a Research Library: Muscat at the Scott Polar Research Institute." Program, 22:1(January 1988); 1-20.

81. *Op. Cit.*

82. Pape, D. L. and Jones R. L. *Op. Cit.*

83. *Op. Cit.*

## Appendix 1. Second-Generation OPAC Features Checklist

### I. USER-SYSTEM INTERACTION/DIALOGUE

1. Choice of dialogue mode ____

2. Search command entry procedure
   a. derived search key ____
   b. formal search language ____
   c. function keys ____
   d. menu selection ____
   e. form filling ____
   f. other ____

3. Flexible, unconstrained dialog structure (enter new search at any point) ____

4. Backup to preceding search state ____

5. Save Search results ____

6. Select session default values (display format, collection, etc.) ____

7. Typing correction/edit procedure ____

8. Interrupt system output ____

9. Stop system processing ____

### II. ONLINE USER ASSISTANCE

1. Training mode/tutorials/lessons (CAI) ____

2. Help command or function key ____

3. Select specific help display by command or menu ____

4. Situation-specific help automatically provided when requested at point of need ____

5. Instructions for correcting typing errors and entering commands/requests ____

6. Current search request displayed on intervening screens ____

7. "How to" prompts for the next step, or search/display options, embedded in the screen dialogue ____

8. Automatically displayed guidance for reformulation/refinement of search ____

## III. SEARCH FORMULATION/REFINEMENT CAPABILITIES

1. Assigned (controlled vocabulary phrase access (includes derived search keys)
   a. author ____
   b. corporate or conference name ____
   c. title ____
   d. series ____
   e. subject ____
   f. classification code ____
   g. other ____

2. Precoordinated, combined-field access
   a. author/title ____
   b. other ____

3. Control/ID number access
   a. LC card number ____
   b. ISBN ____
   c. ISSN ____
   d. Govt. document number ____
   e. CODEN ____
   f. Library call number ____
   g. system ID number (OCLC, RLIN, etc.) ____
   h. other ____

4. Keyword access
   a. author ____
   b. title ____
   c. corporate or conference name ____
   d. subject ____
   e. series ____
   f. notes ____
   g. other ____

5. Boolean search expressions
   a. explicit use of operators ____
   b. implicit (system default action) ____

6. Truncation, word or phrase
   a. explicit use of command symbol ____
   b. implicit (system-supplied) operators ____

7. Limit search results
   a. by field/index specified ______
   b. date ______
   c. language ______
   d. format (*e.g.*, serials, AV, scores) ______
   e. other ______

8. Word proximity search expressions ______

9. Boolean operations on previously created sets (not just current set) ______

10. Post-modify (add term, limit) current search result set ______

11. Special review/browse features
    a. shelf list review ______
    b. browse terms alphabetically near search term ______
    c. browse heading keywords in context ______
    d. cross references displayed ______
    e. browse subject headings/call numbers assigned to documents retrieved by keyword(s) ______
    f. display search history ______

## IV. DISPLAY/PRINT CAPABILITIES

1. Brief, truncated record listing of multiple-hit results ______
2. Single-record search result automatically displayed ______
3. Number of retrievals (*postings*) displayed with index/heading terms ______
4. Full MARC bibliographic information displayable ______
5. Separate data elements in display clearly labeled ______
6. Circulation/availability status information displayed ______
7. Specify part/range of results ______
8. Specify system-defined display format ______
9. Specify data elements for display ______
10. Move backward and forward among results displays ______
11. Sort results for display/print sequence ______
12. Hardcopy print of search results ______

## Appendix 2. Advanced OPACs and Intelligent Retrieval Systems and Software Reviewed or Investigated for the Study

Minnesota State University System/PALS

Dartmouth College Online Catalog

Colorado Alliance of Research Libraries (CARL)

University of Illinois at Urbana-Champaign (LCS/WLN/PC-Interface)

National Library of Medicine (CITE)

Beth Israel Hospital, Boston (Paperchase)

SIRE - Syracuse Information Retrieval Experiment

University of Illinois Engineering Library - Wm. Mischo's Intelligent Microcomputer Front End for Online Literature Searching

General Research Corporation (GRC), (LaserGuide, CD-ROM OPAC)

BiblioFile Intelligent Catalog (CD-ROM)

Bowker/Online Computer Systems, Inc. (CD-ROM)

University of Massachusetts ($I^3R$)

SWALCAP Library Services, UK (LIBERTAS)

Polytechnic University of Central London (OKAPI)

Middlesex Polytechnic University, London (PRECIS-OPAC)

Information Management & Engineering (IME: TINlib)

Etude et Recherche en Linguistique et Informatique (ERLI/RAMEAU)

TOME Associates (TOME*Searchers)

University of Aberdeen, Scotland (KIM)

University of Sheffield, UK (INSTRUCT)

Scott Polar Research Institute, Cambridge University (SPRILIB/Muscat)

Computer Power Group, Canberra, Australia (STATUS with IQ)

Battelle Memorial Research Institute (Vocabulary Switching System)

National Library of Medicine (CANSEARCH)

THE LIBRARY ☆ OF ☆ CONGRESS ☆ ☆