GREENWOOD
LIBRARY
MANAGEMENT
COLLECTION

AUTOMATED INFORMATION RETRIEVAL in LIBRARIES

A Management Handbook

Vicki Anders

Automated Information Retrieval in Libraries

A MANAGEMENT HANDBOOK

Vicki L. Anders

The Greenwood Library Management Collection

GREENWOOD PRESS
New York • Westport, Connecticut • London

Library of Congress Cataloging-in-Publication Data

Anders, Vicki.
Automated information retrieval in libraries : a management handbook / Vicki Anders.
p. cm. — (The Greenwood library management collection, ISSN 0894-2986)
Includes bibliographical references and index.
ISBN 0-313-27361-8 (alk. paper)
1. Information retrieval. 2. Libraries—Automation—Management.
I. Title. II. Series
Z699.A63 1992
025—dc20 91-44401

British Library Cataloguing in Publication Data is available.

Library of Congress Catalog Card Number: 91-44401
ISBN: 0-313-27361-8
ISSN: 0894-2986

First published in 1992

Greenwood Press, 88 Post Road West, Westport, CT 06881
An imprint of Greenwood Publishing Group, Inc.

Printed in the United States of America

The paper used in this book complies with the Permanent Paper Standard issued by the National Information Standards Organization (Z39.48-1984).

10 9 8 7 6 5 4 3 2 1

AUTOMATED INFORMATION RETRIEVAL IN LIBRARIES

Recent Titles in
The Greenwood Library Management Collection

Strategic Marketing for Libraries: A Handbook
Elizabeth J. Wood, with assistance from Victoria L. Young

The Smaller Academic Library: A Management Handbook
Gerard B. McCabe, editor

Operations Handbook for the Small Academic Library
Gerard B. McCabe, editor

Data Bases for Special Libraries: A Strategic Guide to Information Management
Lynda W. Moulton

Time Management Handbook for Librarians
J. Wesley Cochran

Academic Libraries in Urban and Metropolitan Areas: A Management Handbook
Gerard B. McCabe, editor

Managing Institutional Archives: Foundational Principles and Practices
Richard J. Cox

Preface

It was not so many years ago that automated information retrieval meant exclusively online searching by librarian intermediaries. Books and articles on the subject discussed dealing with a handful of online vendors, training librarians to be search analysts, and how to afford it. Searches were performed on tiny Texas Instruments Silent 700 terminals at 300 baud and were printed on thermal paper that curled like an ancient scroll and quickly faded to a uniform and barely legible grey. It was state-of-the-art technology. Graduating to 1,200 baud was astonishing.

End-users entered the picture in the 1980s. Although I had performed ready-reference searches for a couple of years previously, it was the end-users who expanded my contact with automated information retrieval. I was the bibliographic instruction librarian at the time, and spent a good many hours developing slide-tape shows, written manuals and computer-assisted instruction programs to teach end-users how to use BRS/After Dark. I spent many more hours standing at their shoulders explaining which key to punch, the difference between AND and OR, and how to get the computer to print citations. It was the first time I ran into this phenomenon: No matter how many manuals the end-user has read or how many demonstrations the end-user has seen, as soon as the end-user sits down at the keyboard the question comes up, "Now what do I do?" End-users have searched BRS/After Dark and Knowledge Index at Texas A&M University's Evans Library since 1982, and the question is still asked by the majority of new users. But enough have figured it out that end-users now account for 75 percent of the library's total online connect hours. I have been the manager of the library's automated information retrieval service for several years now, and I hear the question every day.

In 1986 Evans Library installed several CD-ROM databases; in 1990 Wilson database tapes were loaded with the library's NOTIS online public access catalog. The automated information retrieval picture has changed considerably since the early days of the 300 baud TI terminal.

This book attempts to illuminate the total picture of an integrated automated information retrieval service by showing how online, ondisk, and locally loaded databases work together to serve search analysts, librarians, and end-users. The book is intended for the managers of the services who deal with the day-to-day details of providing mediated searches, discretionary searches, and end-user searches, or for those managers who are considering adding a new service, and for library management personnel who want a review of the automated information retrieval options and the procedures and costs involved in providing them. Automation of one sort or another consumes more and more of library budgets and library staff each year, and new developments in the technology and techniques of automated information retrieval constantly change the delivery of information services. It is important to understand all the options available, how they affect each other and how they affect the library in general, in order to select the right configuration of options for any one library.

Selective annotated reference sections follow each chapter and should be used as guides to additional reading. I have tried to include books of collected readings, either of reprinted articles or original contributions, because they bring together a variety of viewpoints in one convenient package. Case study articles illustrating unique solutions to particular problems, articles showing a good research design or methodology that can be repeated and tested in a local environment, literature reviews, and surveys that illustrate how a variety of libraries are using and applying automated information retrieval have been selected for inclusion. Very few articles published prior to 1980 are listed unless they are landmark articles much cited in the literature. An American bias is very noticeable. References within the text of the chapter lead the reader to articles about the point being discussed; these articles may not necessarily draw the same conclusion that I have in the text.

I am aware that "compact disc" and "optical disc" use the letter *C* instead of the letter *K*, but I have deliberately chosen to spell the word *DISK*. I have chosen the word "ondisk" to provide a distinction from "online" searching and to refer to the portable databases designed to be mounted on a local microcomputer; thus it includes CD-ROM, optical disks, and laser disks, and can be extended to include those databases distributed on floppy disks and intended to be loaded on the capacious hard disks of microcomputers.

I wish to thank the Faculty Academic Study Program of Texas A&M University for the grant of a Faculty Development Leave allowing me the time to research this book. Thanks are also due to my colleagues in the Reference Division of Evans Library who have lived through many changes and additions to their duties and responsibilities brought on by changes and additions to the automated information retrieval services we offer. They have served as the guinea pigs in many research projects. Special thanks go to Dr. Kathy Jackson, an enthusiastic supporter and promoter of automated information retrieval.

AUTOMATED INFORMATION RETRIEVAL IN LIBRARIES

Contents

Illustrations

1

Automated Information Retrieval: Integrating the Options

WHAT IS AUTOMATED INFORMATION RETRIEVAL

Sitting at his terminal in his crowded office late at night, Professor Smith finds just the reference he needs in the library's OPAC (online public access catalog) in that portion of the OPAC where the commercially produced database INSPEC is loaded. He checks the local holdings and discovers that the journal he wants is not in the library's collection. "Rats!" he says, and although he has done it many times before he has to ask for onscreen help to remember how to place an interlibrary loan request via his office computer. Once the onscreen form is filled in and transmitted by E-mail to the library's interlibrary loan office he sighs and briefly regrets the intolerable bother and delays just to get this one article in his hands. The next morning an interlibrary loan clerk prints out the latest batch of requests waiting on the office computer, and as is the practice of interlibrary loan clerks everywhere, checks them against the library's online catalog just to be sure they are legitimate, and then checks the journal titles against a bibliography of online full-text journals. Professor Smith's journal is there along with a few others. The clerk sends an E-mail message consisting of the appropriate interlibrary loan requests to the Automated Information Retrieval Service Office where it is received by the service coordinator, who notes that there are not so many this morning and decides to do them herself right away. She signs on to the full-text databases and downloads the appropriate articles, signs off, does a little manipulating of the downloaded records to link them to the appropriate interlibrary loan requests and the accounting system, then hits the "send" button. The downloaded articles are now in the memory of the campus mainframe computer linked to several microcomputers and terminals all over campus, including Professor Smith's computer where it waits until he comes in at noon, turns on his computer, and finds the article he requested. Instead of recalling the good old days when he had to walk over to the library, check a card catalog, fill out an interlibrary loan form by hand and carry it to the interlibrary loan office, then

wait for days or weeks for the article to appear, Professor Smith grumbles to himself, "Now, why couldn't the library have gotten this to me last night?"

This is not a scene from the year 2000; it is possible, and it is happening today, including Professor Smith's attitude. Automated information retrieval is defined in this text as computerized access to commercially available databases. Therefore, the locally loaded version of INSPEC and the online database of full-text journals fall within the definition, because both are accessed by computer and both are commercially produced, but the library's online catalog of holdings, being locally produced, does not.

A generation has passed, yet librarians still refer to automated information retrieval as "new" technology, when in fact, as early as 1964 computerized searches in batch mode were available on the MEDLARS database produced by the National Library of Medicine. In the early 1960s, abstracting and indexing services began replacing standard typesetting machines with computerized phototypesetting. The purpose was to produce printed indexes faster and cheaper, but the by-product—a machine-readable tape—had a life of its own. In the early years the machine-readable tape could be batch searched, whereby a number of search requests were loaded at one time and allowed to run, and the resulting printouts were handed over to the requesters some hours or days or even weeks later. There was no interaction with the database; a search strategy could not be modified except as part of another batch to be run later. Faster, more powerful computers and reliable telephone networks created the conditions that made online, interactive searching from remote terminals possible. In 1972 DIALOG began commercial operations with just three databases; by 1975 many academic and research libraries in the United States had instituted automated information retrieval services, and a landmark one million remote searched were performed in that year.

By the early 1980s some abstracting and indexing database producers were making their machine-readable tapes available for lease. Libraries could mount the tapes and searching software on mainframe computers that could be accessed by local users. A locally loaded database cut out the middleman vendors and thus cut out connect time charges. Some leased databases require citation charges to be paid as a royalty to the database producer in addition to the lease charge. Locally loaded databases were and still are an expensive option. Also in the early 1980s, Knowledge Index and BRS/After Dark made several databases available for searching on user-friendly systems by end-users.

In 1986 databases on optical disks and compact disks became widely available. Like their online counterparts, these are large databases run by powerful search engines, but they are moderately user-friendly and thus designed for end-users, and like locally loaded databases they have no connect time charges and no citation charges unless the library imposes them. Although they can be expensive to purchase, the per-search cost of a popular CD-ROM database is quite low compared to the online version. Neufeld and Cornog (1986) and Williams (1977) detail the history of automated information retrieval.

In the early 1990s there are literally thousands of databases available, and some of them are available in all the forms mentioned above: online, online for end-users, locally loaded as a leased tape, or locally loaded as an optical disk. A library may have any combination of these forms depending on the clientele served and the service philosophy of the library. The budget greatly influences how much and what kind of searching will be done, and the organization of the library will affect the organization of the automated information retrieval service. This chapter discusses the options available and describes the typical setting of the different options, their operation, environment, clientele, and use. Finally, the chapter describes the impacts of automated information retrieval services on the library and on each other. First, some definitions are necessary.

What Is a Search?

A search is defined as the access to a database or databases necessary to answer a question. A patron wanting the full bibliographic details for an ERIC document so far identified only as ED123 456 can be answered by accessing the ERIC database and typing the character string ED123 456, and printing the single retrieval in bibliographic format. If the online version of ERIC is searched, the cost is a couple of dollars and a couple of minutes for the answer to the patron's question. Many searches are far more complex and require the formulation of a *search strategy*, a formalized model of the question written out in the style that the computerized database can accept. Several sources describe in detail the development of search strategies and the mechanics of online searching, including Gilreath (1984); Li (1985); Armstrong and Large (1988); and Vigil (1988). These should be used with caution because any guide to searching is out of date as soon as it is published, due to the improvements and upgrades the vendors add to their search engines.

For example, a veterinary medicine student needs to find information on the nutritional requirements of elderly cats. The *search analyst*, a librarian with special training in automated information retrieval, explains which databases are most likely to contain citations on the subject, and the patron selects two of them: the Commonwealth Agricultural Bureaux database, which includes Nutrition Abstracts and Index Veterinarius, and BIOSIS. Both databases are widely available through different vendors, and the search analyst decides to use the BRS system to compile a bibliography of sources. The first step is to analyze the question for the concepts it contains, like so:

CATS ELDERLY NUTRITIONAL

It has been decided to search two different databases, and the search analyst will therefore use *free-text* terms rather than the controlled vocabulary particular to a single database. Free-text searching finds words as they occur in any basic index field of a bibliographic citation—title or abstract or descriptors—whereas

controlled vocabulary searching searches only in the descriptor field, and the descriptors are seldom the same in two different databases. The search analyst and the student requesting the search must put their heads together to make a list of synonyms for each of the concept groups, because different authors may use different words to describe the same thing. Author Smith may use "cats" and author Brown may use "felines." The search strategy expands to look like this:

CATS	ELDERLY	NUTRITIONAL
cat	old	nutrition
feline	older	diet
felines	geriatric	dietary
felis	mature	food
felidae	senile	feed
	aged	

Notice that singular and plural forms of words are listed as synonyms because computerized searching is quite literal and searches for character strings exactly as they are entered. Typing "cat" at the terminal keyboard will not normally retrieve "cats." However, the search analyst applies his or her special knowledge of automated information retrieval techniques to revise the list. For example, BRS has an automatic plurals option which, if it has been turned on, will retrieve "cats" when "cat" is typed. In order to avoid typing both "old" and "older", the search analyst revises the character string to read "old$2", using the BRS truncation technique to instruct the computer to search for the character string "old" plus up to two more characters if they exist. The search analyst could apply the same technique to the word stem "feli" by typing "feli$3" and thus avoid consuming online time by typing four variations based on the same word stem. The search analyst points out a possible danger here: searching for the scientific terms felis/felidae can retrieve information on lions, tigers, and other large undomesticated predators, and the patron has made it clear that his research relates to domestic house cats. However, in combination with the other concept groups, any such retrieval would probably deal with the nutrition or diet of elderly tigers, and the patron thinks there might be information of interest to him in such an article. They agree that it is better to leave the scientific terms in and possibly retrieve some citations outside the area of interest, rather than leave them out and possibly miss some relevant citations because the search strategy was too narrowly defined. The search strategy now looks like this:

CAT	ELDER$2	NUTRITIO$3
feli$3	old$2	diet$3
	geriatric	food
	mature	feed
	senile	
	aged	

Further discussion reveals that the patron wants only English language articles and only publications from the last ten years. The search analyst makes the appropriate notations on the search strategy worksheet after consulting the BRS AidPages to see how the two different databases handle language and year limits. The search analyst explains how logical and positional operators such as AND, OR, NEAR, and WITH will be applied to the concept groups, and asks once again if there are any other terms the patron wants included or excluded from the search. There are not, and the search analyst says, "Let's go online"—the rallying cry of all enthusiastic search analysts.

From the terminal or microcomputer in the library, the search analyst places a long-distance call through a local packet-switching network such as Sprintnet or Tymnet to the BRS computer in Chicago, enters an account password and a security password, and then instructs the computer to open the appropriate database. The search strategy is typed as it was worked out on the search strategy worksheet, like so:

```
ENTER DATABASE NAME_:  caba

*SIGN ON        9:50:42          06/20/91

CABA 1972-MAY 1991 (9105)

BRS SEARCH MODE - ENTER QUERY

     1_:  feli$3

 RESULT          7620 DOCUMENTS

     2_:  old$2 or geriatric or mature or senile or aged

 RESULT        175994 DOCUMENTS

     3_:  diet$3 or food or feed

 RESULT        333882 DOCUMENTS

     4_:  1 and 2 and 3

 RESULT            47 DOCUMENTS
```

At this point, the search analyst prints the titles and descriptors of a few of the citations retrieved so that the patron can examine them to see if the search is retrieving the type of information needed. If not, the search can be refined by adding or subtracting terms or limiting terms to certain fields. Once the patron is satisfied with the samples retrieved, the full set can be printed either online (citations printed at the library terminal immediately) or offline (citations printed at the BRS site and sent to the library by mail). The search analyst then saves (stores temporarily in the BRS computer's memory) the basic search strategy, changes to the BIOSIS database, and executes the saved search. Once again the

search strategy can be modified online if the sampled citations are not precisely what the patron wants. After the second set of prints has been ordered, the search analyst signs off and disconnects from the BRS computer and the telephone network.

This is one search. The same search could be performed by an end-user through the BRS/After Dark system, or it could be performed on the compact disk versions of the two databases. The process of accessing a database or databases to find the answer to a question constitutes a search.

Search Engine or Search Software or Search Protocols

In the sample search above, the BRS truncation symbol—$—and the positional operators such as NEAR or WITH, are examples of the commands that must be used when searching a BRS database. The truncation symbol used by a different vendor, such as DIALOG's question mark, cannot be used as the truncation symbol when searching BRS. SilverPlatter uses a different set of commands, as does ORBIT, as does UMI. Each vendor has a different set of commands that must be used when searching that vendor's databases, otherwise the system will not work properly. For example, if one uses the DIALOG truncation symbol while searching in a BRS database (such as 1- geriatric or aged or senile or mature or old?), the BRS computer interprets the ? entered at the end of the character string as the command to erase the entire line, which it does. It presents the searcher with the old prompt (1-). The searcher must reenter the character string and this time remember to use the correct truncation command, a very frustrating experience which has happened to practically every searcher. The truncation symbol is an example of a *search protocol*, an abbreviated command instructing the host computer to perform a particular function. *Search software* usually refers to the physical machine-readable tape or diskette that contains the program allowing a computer to perform the functions. *Search engine* refers to the system of commands and techniques and user interfaces unique to a particular vendor's products. In practice, the three phrases are used interchangeably as shorthand to say, "If one searches a SilverPlatter database, one must use SilverPlatter commands." The proliferation of search engines complicates the lives of searchers as it becomes more and more difficult to remember which system uses ? or $ or : or # as the truncation symbol, and which ones have automatic plurals.

Databases

Technically, just about any accumulation of alpha or numeric or symbolic characters can be a database. An address book or a recipe file can be considered a database. If the address book or recipe file is loaded on a floppy disk using database management software (DBMS; also called database management

system), it becomes a machine-readable file. In the context of this book, a *database* is machine-readable, computerized, automated, and has an associated search engine allowing the data to be retrieved, sorted, or otherwise manipulated, and is commercially produced by a publisher, association, government agency, or other organization with the intention of disseminating the information publicly, usually for a price. The last point—commercial production and distribution of the database—distinguishes the subject of this book from the many OPACs, locally produced ready-reference files, and computerized address books being compiled. There are several different types of databases:

1. *Bibliographic*—usually the machine-readable equivalent of an abstracting or indexing tool. A bibliographic database provides surrogate information, that is, citations to works such as articles or monographs rather than the text of works. These are the databases that have been available the longest and are used most often in libraries. ERIC is an example of a bibliographic database, produced by the Educational Resources Information Center and corresponding to the printed tools *Current Index to Journals in Education* and *Resources in Education*. A typical entry in ERIC and other bibliographic databases consists of a unique accession number; author, title, and publication information; descriptors or subject headings usually assigned as a controlled vocabulary; and an abstract, which in some databases is quite lengthy. Some databases may have coded fields to aid retrieval, such as the Biosystematic Codes in the BIOSIS (Biological Abstracts) database, or the event codes and product codes in the PTS (Predicasts Terminal System) databases. A bibliographic database can be quite small, containing only a few thousand entries, or so massive with millions of entries that it would be unwieldy to search and so has been split into smaller databases based on ranges of publication years. BIOSIS, CA Search (Chemical Abstracts), and MEDLINE (Index Medicus) are examples of these massive databases. Some bibliographic databases do not directly correspond to a single printed tool but combine two or more printed equivalents in one database. ERIC is an example; INSPEC is another, combining *Physics Abstracts, Electrical and Electronics Abstracts*, and *Computer and Control Abstracts*. Some bibliographic databases are available only in machine-readable form and do not have print equivalents, such as InfoTrac and other databases from IAC (Information Access Company).

2. *Full-Text*—the entire document, possibly excluding illustrations, stored in machine-readable form. The widespread use of computerized phototypesetting to produce printed journals, newspapers, and books also produces a machine-readable tape. Therefore, the possibilities for producing full-text databases are enormous, but not yet fully realized. Many major newspapers and a small percentage of journals are available in full-text. The first publicly available full-text files were the law databases provided by LEXIS and WESTLAW. Many reference tools such as the *Merck Index* and *Encyclopedia of Associations* are