

I THE SEARCH FOR INFORMATION IN THE ONLINE AGE

This chapter identifies some of the problems facing the information-seeker in a world of information overload. It introduces the concept of online searching and attempts to show the usefulness of such searching by giving examples of the wide variety of information-bearing materials that are available online today. The chapter then explains the basic mechanics of information retrieval, which apply equally to both manual and computerized information systems of all types—library catalogs, CD-ROMs, online database systems, and the Internet.

Information Overload

The problems of organizing information in order to find it when it is required are nothing new. From as early as the seventh century BC, collections of information-bearing artifacts existed in the form of stone tablets, and later parchment scrolls. However, for many centuries, the emphasis was on preservation of the cultural heritage of society in an uncertain world, rather than on its retrieval in response to a request.¹ Documents were relatively rare, and so few people could read that there was little call for access to individual items. Today, the situation has totally changed. The pressures brought about by the explosion in the amount of material being published annually over the last 50 years have highlighted the problems involved in retrieving a single desired item from the proverbial “haystack” of available information. The focus has moved from collection and preservation to retrieval and selection.

It is all too easy to assume that modern technology can solve this retrieval problem, or at least simplify the process. The use of computers and communication technologies has had an enormous influence on the way that information is produced, organized, stored, searched, and transmitted, and has certainly made more information accessible to more people. But these new technologies have failed to solve the intellectual problems of information retrieval. In fact, in some ways they have made retrieval more difficult. On the one hand, technology has speeded the search process, increased the amount of information that can be accessed by a single search, and enabled the searcher to use much more sophisticated search strategies. On the other hand, the amount of information available is daunting, and the variety of different ways to access it means that selection and evaluation are key ingredients to successful searching. Thus, technological developments have resulted in an increased need for user training if the most efficient use is to be made of these powerful new retrieval systems.

What Information Can You Find Online?

The heart of an online retrieval system is the information it contains—the files or databases that are available for searching. Databases come in all shapes and sizes, ranging from vast files such as MEDLINE or CA Search (containing around 9 million and 12 million records respectively), to tiny specialized files such as the Philosopher's Index (just under 200,000 records) or AIDSLINE (144,000 records, but growing fast). Although these are examples of bibliographic files, there is now a wide range in types of information available for searching online, much of it in full text. Different databases not only contain different types of information but also are useful for a variety of purposes, ranging from quick verification of a reference to factual data to in-depth research. Today it is possible to find answers to almost any type of reference query using online resources.

In this book we are going to concentrate on one of these online search services, the DIALOG system (now owned by M.A.I.D. plc). The search language that you will learn will be the DIALOG command language, and the majority of our examples are taken from the DIALOG system, though examples are also given from other systems. The reason for this concentration on DIALOG is our belief that it is very difficult to learn more than one new language at a time, and that it is better to become fluent in one new language before trying another. We also believe that, because most online systems have a very similar range of commands, once one understands the kinds of features available and becomes proficient on one system, it is not difficult to transfer that knowledge to another system with a slightly different command language. Both CD-ROMs and the WWW have attempted to make their systems easier to use by the uninitiated, but better searching on these systems will also result from a general understanding of what the computer is doing in any given situation.

It is also true that the DIALOG command language is among the most powerful, full-featured (and therefore complex) language. Learning it and what it can do will serve you well in learning, using, and evaluating other systems, including Web-based ones. —JWJ

Some search services provide information in a single subject field. For example, Mead Data Central has specialized in legal information, with two databases covering federal statutes and case law. DIALOG, however, is a multisubject system, so that it provides a particularly useful example to illustrate the diversity of information that can be accessed online.

There are four basic types of databases currently available for online access:

- Bibliographic citations (sometimes including abstracts)
- Full-text documents
- Directory sources
- Numeric data

It is not uncommon for a single search subject to be covered by several databases, which will often provide different types of information. The degree of overlap between files and the occurrence of unique items will vary considerably by both subject field and database. In order to demonstrate the diversity of information available, let us look at how the records in different files provide information suitable for answering a variety of requests.

Bibliographic Information

Although the original use of online systems was normally to produce a bibliography of citations in response to a subject search request, the same type of record is also useful for checking an incomplete or misremembered reference. For example, a request for "that book about tough California writers" can be swiftly identified by using a database such as the Library of Congress MARC file or the online version of R. R. Bowker's *Books in Print*®. The resulting citation in *Books in Print*® (fig. 1.1) contains information not unlike that in a standard catalog record.

Fig. 1.1. Books in Print® record.

```
00598185 1034311XX STATUS: Out of print (06-92)
TITLE: California Writers: Jack London, John Steinbeck, the Tough Guys
AUTHOR: Martin, Stoddard
PUBLISHER: St Martin PUBLICATION DATE: 01/1984 (840101)
NO. OF PAGES: 224p.
LCCN: 82-020451
BINDING: Trade - $22.50
ISBN: 0-312-11420-6
VOLUME(S): N/A
ORDER NO.: N/A
IMPRINT: N/A
STATUS IN FILE: New (84-02)

LIBRARY OF CONGRESS SUBJECT HEADINGS: LONDON, JACK, 1876-1916 (00280707)
```

The same record on LC MARC (fig. 1.2) looks rather more complicated.

Fig. 1.2. LC MARC record.

```
1637993 LCCN: 82020451
California writers ; Jack London, John Steinbeck, the tough guys /
Stoddard Martin
Martin, Stoddard, 1948-

New York : St. Martin's Press, viii, 224 p. ; 23 cm.
PUBLICATION DATE(S) : 1983
PLACE OF PUBLICATION: New York
ISBN: 0312114206
LC CALL NO.: PS283.C2 M35 1983 DEWEY CALL NO.: 813/.54/099794
RECORD STATUS: Increase in encoding level from prepublication record
BIBLIOGRAPHIC LEVEL: Monograph
LANGUAGE: English
GEOGRAPHIC LOCATION: California
NOTES:
Includes bibliographical references and index.
DESCRIPTORS:
American fiction -- California -- History and criticism; American fiction
20th century -- History and criticism; California in literature
```

Here are some other examples of bibliographic records to prove that even a single type of record can vary greatly (see figs. 1.3 and 1.4, p. 4).

Fig. 1.3. Magazine Index record.

12535757 DIALOG File 47: MAGAZINE INDEX *Use Format 9 for FULL TEXT*
'Hearty' vitamins; sparing arteries with megadose supplements. (research
using soybean-oil and vitamin E)
Raloff, Janet
Science News v142 p76(1) August 1, 1992
SOURCE FILE: MI File 47
CODEN: SCNEB ISSN: 0036-8423
AVAILABILITY: FULL TEXT Online LINE COUNT: 00092
ABSTRACT: Scientists are conducting tests that boost LDL oxidant defensive
benefits in order to discourage plaque buildup in the human body. The
study includes large doses of soybean-oil capsules and vitamin E.
Researchers say that vitamin therapy will probably be used to ward off
heart disease.
DESCRIPTORS: Atherosclerosis--Prevention; Cholesterol, LDL--Research;
Oxidation--Research; Vitamin therapy--Research

Fig. 1.4. BIOSIS Previews record.

9610212 BIOSIS Number: 94115212
HEPATIC CYTOTOXICITY AND MUTAGENIC POTENTIAL OF SODIUM SELENITE VITAMIN E
ICCF BUCURESTI ON MOUSE BONE MARROW IN-VIVO
SOCACIU C; PASCA I; LISOVSCI C
USACN, STR. MANASTUR NR. 3, CLUJ-NAPOCA 3400, ROMANIA.
BUL INST AGRON CLUJ-NAPOCA SER ZOOTEH MED VET 46 (0). 1992. 121-127.
CODEN: BIAVD
Full Journal Title: Buletinul Institutului Agronomic Cluj-Napoca Seria
Zootehnie si Medicina Veterinara
Language: ROMANIAN
Subfile: BA (Biological Abstracts)
The cytotoxic and mutagenic potential of a Natrium Selenite - vitamine E
mixture through "in vivo" treatment was investigated. Swiss-Albino male
mice were orally administered with 0.7, 1.4 or 3.5 mg drug/kg body weight
and sacrificed after 24 hrs. The micronucleus test was applied on bone
marrow cells and the percentage of micronucleated polychromatic
erythrocytes was significantly increased versus control only for 3.5 mg
drug/kg b. w. This effect was comparable with 550 mg/kg cyclophosphamide
(positive control) treatment in the same conditions. The hepatic
cytotoxicity was investigated in the same experimental and control groups
through evaluation of some marker enzymes activities: Ca-ATPase. G-6-Pase,
lactatedehydrogenase (LDH) sorbitolydehydrogenase (SDH), peroxidase (Px)
and also through nucleotide protein ratio (DO260/280). All drug
concentrations revealed a liver metabolic activation, significantly
increased versus control. The experiment revealed a mutagenic and a
cytotoxic potential for Natrium Selenite--Vitamin E mixture especially
around 3.5 mg/kg but for the practical significance of this result we must
consider that the therapeutic dose is always much more inferior to the
toxicity limit.
Descriptors/Keywords: ERYTHROCYTE
Concept Codes:
*02506 Cytology and Cytochemistry-Animal
*03506 Genetics and Cytogenetics-Animal
*10063 Biochemical Studies-Vitamins
*10066 Biochemical Studies-Lipids
*13016 Metabolism-Fat-Soluble Vitamins
*14006 Digestive System-Pathology
*15008 Blood, Blood-Forming Organs and Body Fluids-Lymphatic Tissue and
Reticuloendothelial System
*18004 Bones, Joints, Fasciae, Connective and Adipose Tissue-Physiology
and Biochemistry
*22504 Toxicology-Pharmacological Toxicology (1972-)
Biosystematic Codes:
86375 Muridae
Super Taxa:
Animals; Chordates; Vertebrates; Nonhuman Vertebrates; Mammals; Nonhuman
Mammals; Rodents

As you can see, the amount of information in the records differs vastly, and they are obviously aimed at very different audiences. Such differences in record content and length affect the ways in which you can search the files, given that any part, or field, of the record is usually searchable.

Full-Text Records

In some databases each record contains the entire text of a document, and online systems are able to search for the occurrence of any single word or phrase in these full-text documents. Many newspapers, legal cases, and reference sources, such as encyclopedias and directories, are now available in full-text format. Full-text is also widely available on the Web, ranging from whole journal articles or even books (e.g., many of the classics that are out of copyright) to personal information and even advertising. Figure 1.5 provides an example from an international newspaper article. Many of these full-text records are very long, so they require special strategies when searching to zero in on the exact information needed and to avoid wasting a lot of time and money.

Fig. 1.5. London Times record (extract).

09544138
 OJ CHALLENGED TO CONFESS;MURDER
 Times of London (TL) - Thursday, February 13, 1997
 By: Giles Whittell
 Section: Overseas news
 Word Count: 193

TEXT:
 FRED GOLDMAN, whose son was murdered in 1994 at the same time as Nicole Brown Simpson, O.J. Simpson's former wife, has issued a challenge to Mr Simpson, urging him to publish a detailed signed confession to the murders.

In return, Mr Goldman said he would make no attempt to collect the \$21 million (Pounds 12.8 million) in damages awarded to him this week against Mr Simpson by a civil jury, which held the former football star responsible for the two deaths.


It was not clear yesterday whether the Brown family would offer to drop their claim to \$12.5 million in punitive damages, and Mr Simpson has yet to respond to the challenge. Legal experts pointed out, however, that he was unlikely to accept, not least because that could expose him to perjury charges.

Mr Goldman first issued his challenge on Tuesday on Salem Radio Network and was asked by Mark Gilman on the Alan Keyes Show if it was that simple. "Easy to say, easy to do, never going to happen," Mr Goldman said. "This person hasn't owned responsibility for any of his actions through his lifetime."

* Race to replace O.J., page 19.

Full-text information on the Web ranges from similar informative records (see fig. 1.6, p. 6), to blatant commercial advertising (see fig. 1.7, p. 7), and it is worth pointing out that the Web is a resource that is unedited and largely unsupervised, so extra careful evaluation of retrieved material is even more essential than with bibliographic records.

Fig. 1.6. Internet full-text journal record (abbreviated).



[Eating](#) | [Drinking](#) | [Playing](#) | [Bon Appétit](#) | [Gourmet](#) | [Forums](#)
[Home](#) | [Text-Only Index](#) | [Go to Epicurious Travel](#)

epicurious
RECIPE FILE

[Click here to email a copy of this recipe to a friend](#)

INDONESIAN-STYLE GRILLED EGGPLANT WITH SPICY PEANUT SAUCE WHITE

1 eggplant (about 1 1/4 pounds), cut into 1/2-inch-thick slices
1 garlic clove, minced
1 shallot, minced
a 2-inch-long fresh hot red chili, chopped fine (wear rubber gloves), or 1/4 teaspoon crushed red pepper flakes
2 teaspoons Oriental sesame oil
1/4 cup ground roasted peanuts
2 teaspoons soy sauce
1 teaspoon sugar
2 teaspoons fresh lemon juice, or to taste
vegetable oil for brushing the eggplant

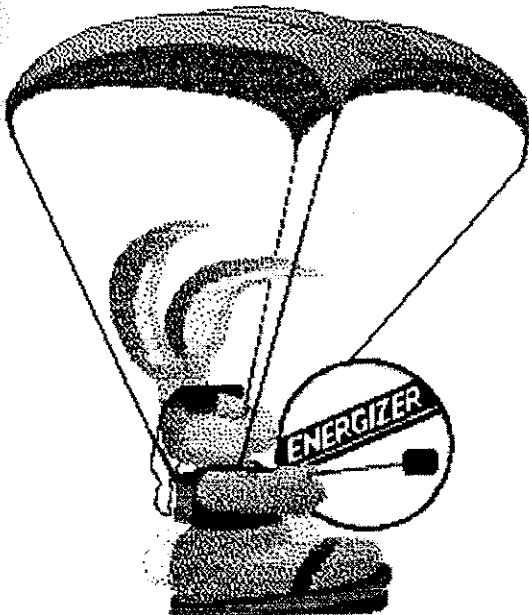
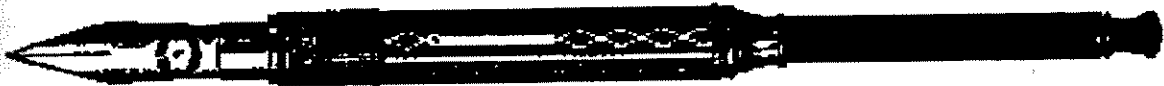
Sprinkle the eggplant lightly with salt, let it drain in a colander for 1 hour, and pat it dry. In a small saucepan cook the garlic, the shallot, and the chili in the sesame oil over moderately low heat, stirring occasionally, until the vegetables are softened, add the peanuts, and cook the mixture, stirring, for 1 minute. Add the soy sauce, the sugar, the lemon juice, and 1 cup water, boil the mixture, stirring occasionally, until it is thickened slightly, and add salt and pepper to taste. Brush the eggplant, patted dry, with the vegetable oil and grill it on an oiled rack set 5 to 6 inches over glowing coals, turning it, for 7 to 8 minutes, or until it is just cooked. Transfer the eggplant to a serving plate and spoon the peanut sauce over it.

Serves 4.

Gourmet
July 1990
Debbie White: Atlanta, Georgia

Fig. 1.7. Internet advertisement record.

Energizer Bunny



It's the famous Energizer Bunny, star of award-winning TV commercials, now appearing on your Mac. Yes! You heard it right. This bunny with an attitude will invade your Mac with cool desktop wallpaper and sizzling hot screen savers. But no! Energizer Bunny doesn't stop there. Just when you least expect it, he proudly parades across your screen beating his drum for everyone to hear. Amaze your friends as the original Energizer Bunny struts his stuff with live-action sound and full animation. Don't wait, don't hesitate! Be the first to have the one and only Energizer Bunny show off on your Mac. He'll dazzle and impress, as he keeps going and going and going...

Requires Mac with 1.44MB floppy, running System 6.0.7 or later, 2MB of RAM, color or gray-scale monitor.

To order a discounted copy of Energizer Bunny (list price: \$29.95), consult the chart below for pricing information. There are no hidden charges. I pay taxes where appropriate and ship all orders the same day they are received via First Class Mail service within the United States and Air Mail delivery elsewhere.

Directory Information

Directory files are systematically arranged listings of people, organizations, or institutions, providing addresses, affiliations, qualifications, and so forth. These are used to locate experts or organizations through their addresses, phone numbers, zip codes, specialization, and so on. This kind of information is often among the most popular quick reference types of query in libraries, particularly public libraries. Figures 1.8 (*Marquis Who's Who*) and 1.9 (*American Library Directory*) are examples of this type of directory resource that are available online. (See p. 8.)

Fig. 1.8. Marquis Who's Who record.

00934100 Record provided by: Marquis
Perot, H. Ross
OCCUPATION(S): investments and real estate group executive; data
BORN: 1930
SEX: Male
FAMILY: married; 4 children.
EDUCATION:
Ed., U.S. Naval Acad.
CAREER:
founder, Perot Systems Corp., Washington, 1988-
now with, The Perot Group, Dallas
chmn., chief exec. officer, also dir., Electronic Data Systems Corp.,
Dallas, to 1986
founder, Electronic Data Systems Corp., Dallas, 1962-84
data processing salesman, IBM Corp., 1957-62
MILITARY:
Served with USN, 1953-57
AWARDS:
Recipient Internat. Disting. Entrepreneur award, U. Man., 1988.

Office: Dallas, TX

Fig. 1.9. American Library Directory record.

00021412 0733532000 LIBRARY RECORD
OFFICIAL NAME: SYRACUSE UNIVERSITY LIBRARY - E S Bird Library
LIBRARY TYPE: COLLEGE & UNIVERSITY
ADDRESS: 222 Waverly Ave
Syracuse, NY
13244-2010
SAN (Standard Address Number): 354-2645
TELEPHONE NUMBER(S): 315-443-2573; Interlibrary Loan Service Tel. No.:
443-3725

LIBRARY HOLDINGS:
BK VOLS: 2,650,995 PER SUB: 9015 MICRO: Total 3,238,652

SPECIAL COLLECTIONS: Stephen Crane First Editions & Manuscripts; Spire
Collection on Loyalists in the American; Revolution; Novotny Library of
Economic History; William Hobart-Royce Balzac Coll; Library; Sol
Feinstone Library; Leopold Von Ranke Library; Shaker Coll; Oneida
Community Coll; Margaret Bourke-White Coll; Marcel Breuer Coll; Peggy
Bacon Papers; Anna Hyatt Huntington Papers; Earl R Browder Papers;
Rudyard Kipling First Editions; Cartoonist Coll; Science Fiction Books &
Manuscripts; Modern American Private Press Books; Gerrit & Peter Smith
Coll; C P Huntington Papers; Averill Harriman Gubernatorial Papers;
Dorothy Thompson Papers; Street & Smith Archive; Arna Bontemps Papers;
Grove Press Archive; Continuing Education Coll; Benjamin Spock Papers;
Mary Walker Papers; Albert Schweitzer Papers; William Safire Coll
US DOC DEP STATE DOC DEP

Numeric Data

Databases containing numeric information of various types are also available for on-line searching. A file such as Donnelly Demographics can provide selected information from the most recent U.S. census, enhanced with estimates of the current situation and even offering five-year projections for certain categories of data. Figure 1.10 (pp. 9-11) presents a very abbreviated section of a single record, but it gives some idea of the wealth of data that can be found in numeric files. It is particularly useful to be able to retrieve this type of information in machine-readable form because the data can then be uploaded to a statistical or database program for analysis and manipulation.

Fig. 1.10. Donnelly Demographics record (extract).

00007912
 ST JOHNSVILLE (Zip Code 13452)

Level: ZIP
 State: NY (New York)
 County: MONTGOMERY
 SMSA: ALBANY-SCHENECTADY-TROY NY (0160)
 PMSA/MSA: ALBANY-SCHENECTADY-TROY, NY MSA (0160)
 ADI: ALBANY-SCHENECTADY-TROY
 DMA: ALBANY-SCHENECTADY-TRO
 SAMI: ALBANY
 Zip Code: 13452 (ST JOHNSVILLE, NY)
 City or Place: ST JOHNSVILLE

Totals & Medians	1980 Census	1991 Estimate	% Change 80 to 91	1996 Projection
Total Population	4,766	4,494	-5.7	4,358
Total Households	1,647	1,570	-4.7	1,529
Household Population	4,763	4,487	-5.8	4,351
Average Household Size	2.9	2.9	-1.2	2.8
Average Household Inc.	\$14,254	\$28,152	97.5	\$34,720
Median Household Income	\$13,019	\$23,833	83.1	\$29,070

Population by Age and Sex

Population by Age	1980 Census		1991 Estimate		1996 Projection	
	Number	Pct.	Number	Pct.	Number	Pct.
Total	4,766	100.0%	4,494	100.0%	4,358	100.0%
0 - 4	362	7.6%	374	8.3%	356	8.2%
5 - 9	414	8.7%	352	7.8%	351	8.1%
10 - 14	421	8.8%	321	7.1%	331	7.6%
15 - 19	454	9.5%	347	7.7%	302	6.9%
20 - 24	326	6.8%	367	8.2%	324	7.4%
25 - 29	334	7.0%	390	8.7%	344	7.9%
30 - 34	299	6.3%	297	6.6%	363	8.3%
35 - 39	285	6.0%	287	6.4%	278	6.4%
40 - 44	216	4.5%	266	5.9%	267	6.1%

10 / 1—The Search for Information in the Online Age

Fig. 1.10. Donnelly Demographics record (continued).

45 - 49	206	4.3%	237	5.3%	246	5.6%
50 - 54	259	5.4%	189	4.2%	219	5.0%
55 - 59	300	6.3%	169	3.8%	168	3.9%
60 - 64	270	5.7%	196	4.4%	150	3.4%
65 - 69	225	4.7%	221	4.9%	168	3.9%
70 - 74	146	3.1%	191	4.3%	178	4.1%
75 - 79	102	2.1%	142	3.2%	144	3.3%
80 - 84	73	1.5%	79	1.8%	97	2.2%
85 +	73	1.5%	69	1.5%	74	1.7%
< 15	1,197	25.1%	1,047	23.3%	1,038	23.8%
65 +	619	13.0%	702	15.6%	661	15.2%
75 +	248	5.2%	290	6.5%	315	7.2%
Median Age	31.2		31.6		32.4	
Median Age Adult Pop.	45.1		42.6		42.6	

Female Population by Age

Total	2,464	100.0%	2,322	100.0%	2,253	100.0%
0 - 4	171	6.9%	182	7.8%	174	7.7%
5 - 9	223	9.1%	172	7.4%	171	7.6%
10 - 14	200	8.1%	152	6.5%	162	7.2%
15 - 19	239	9.7%	183	7.9%	143	6.3%
20 - 24	170	6.9%	178	7.7%	172	7.6%
25 - 29	161	6.5%	202	8.7%	167	7.4%
30 - 34	166	6.7%	156	6.7%	189	8.4%
35 - 39	136	5.5%	142	6.1%	146	6.5%
40 - 44	108	4.4%	142	6.1%	132	5.9%
45 - 49	110	4.5%	119	5.1%	132	5.9%
50 - 54	128	5.2%	95	4.1%	110	4.9%
55 - 59	165	6.7%	91	3.9%	85	3.8%
60 - 64	137	5.6%	101	4.3%	82	3.6%
65 - 69	116	4.7%	124	5.3%	89	4.0%
70 - 74	74	3.0%	105	4.5%	104	4.6%
75 - 79	66	2.7%	83	3.6%	84	3.7%
80 - 84	48	1.9%	46	2.0%	61	2.7%
85 +	46	1.9%	49	2.1%	50	2.2%
< 15	594	24.1%	506	21.8%	507	22.5%
65 +	350	14.2%	407	17.5%	388	17.2%
75 +	160	6.5%	178	7.7%	195	8.7%
Median Age	32.0		32.9		33.6	
Median Age Adult Pop.	46.2		43.6		43.7	

Male Population by Age

Total	2,301	100.0%	2,172	100.0%	2,107	100.0%
0 - 4	191	8.3%	192	8.8%	182	8.6%
5 - 9	191	8.3%	180	8.3%	180	8.5%
10 - 14	221	9.6%	169	7.8%	169	8.0%
15 - 19	215	9.3%	164	7.6%	159	7.5%
20 - 24	156	6.8%	189	8.7%	152	7.2%
25 - 29	173	7.5%	188	8.7%	177	8.4%
30 - 34	133	5.8%	141	6.5%	174	8.3%
35 - 39	149	6.5%	145	6.7%	132	6.3%

Fig. 1.10. Donnelly Demographics record (continued).

40 - 44	108	4.7%	124	5.7%	135	6.4%
45 - 49	96	4.2%	118	5.4%	114	5.4%
50 - 54	131	5.7%	94	4.3%	109	5.2%
55 - 59	135	5.9%	78	3.6%	83	3.9%
60 - 64	133	5.8%	95	4.4%	68	3.2%
65 - 69	109	4.7%	97	4.5%	79	3.7%
70 - 74	72	3.1%	86	4.0%	74	3.5%
75 - 79	36	1.6%	59	2.7%	60	2.8%
80 - 84	25	1.1%	33	1.5%	36	1.7%
85 +	27	1.2%	20	0.9%	24	1.1%
< 15	603	26.2%	541	24.9%	531	25.2%
65 +	269	11.7%	295	13.6%	273	13.0%
75 +	88	3.8%	112	5.2%	120	5.7%
Median Age	30.1		30.1		31.0	
Median Age Adult Pop.	44.0		41.5		41.5	

Population by Race

Total	4,766	100.0%	4,494	100.0%	4,358	100.0%
White	4,739	99.4%	4,449	99.0%	4,304	98.8%
Black	2	0.0%	7	0.2%	8	0.2%
Other	24	0.5%	38	0.8%	46	1.1%
Hispanic	20	0.4%	21	0.5%	22	0.5%

	1980 Census		1991 Estimate		1996 Projection	
	Number	Pct.	Number	Pct.	Number	Pct.
Household Income						
\$ 0- 7,499	399	24.2%	181	11.5%	145	9.5%
\$ 7,500-14,999	569	34.5%	270	17.2%	197	12.9%
\$15,000-24,999	507	30.8%	377	24.0%	306	20.0%
\$25,000-34,999	120	7.3%	319	20.3%	285	18.6%
\$35,000-49,999	38	2.3%	256	16.3%	292	19.1%
\$50,000-74,999	13	0.8%	123	7.8%	200	13.1%
\$75,000 +	1	0.1%	44	2.8%	104	6.8%

1991 Estimate

Neighborhood Mobility

Household Moved In:

Most Recent Year	179
Last 5 Years	539
6 - 9 Years Ago	221
10 - 14 Years Ago	226
15+ Years Ago	528

SocioEconomic Status:

Socioeconomic Status Score	27
Private Sector Employment	1,295

Other numeric databases are of particular interest in the business environment, such as the various U.S. and international financial and company files. For example, information regarding U.S. import and export figures is available from a file such as Piers Exports (see fig. 1.11). It is also easy to collect data on the financial status of a particular U.S. or international public corporation using one of the Moody files. Figure 1.12 provides the record for the IBM Corporation from Moody's Corporate Profiles.

Fig. 1.11. Piers Exports record.

08532492
 Product Exported: VOLKSWAGEN 85 JETTA
 Product Code: 6921000 (AUTOMOBILES, MOTOR VEHICLES, TRAILERS)
 Weight of Cargo: 1900 POUNDS
 Number of Units of Cargo: 1 UNITS

Date of Shipment (YY/MM/DD): 920831

U.S.-Based Exporter: A & M INTL SERVICE
 Company Location: MIAMI, FL

U.S. Port of Loading: PT EVERGLADES (5203)

Destination Point: CALLAO (33303), PERU (333)

Fig. 1.12. Moody's Corporate Profiles record (abbreviated).

0004427
 INTERNATIONAL BUSINESS MACHINES CORP.

MOODY'S NUMBER: 00004427
 DUNS NUMBER: 00-136-8083

STATISTICAL RECORD				
12/31/91	12/31/90	12/31/89	12/31/88	12/31/87
Operating profit margin%				
1.5	16.0	11.0	14.7	14.3
Book value				
56.96	67.79	61.28	62.23	60.09
Return on equity%				
NIL	14.1	9.8	13.9	13.7
Return on assets%				
NIL	6.9	4.8	7.5	8.3
Average yield%				
4.3	4.4	4.2	3.8	3.2
P/E ratio-high				
NIL	11.7	20.2	14.0	20.2
P/E ratio-low				
NIL	9.0	14.4	11.2	11.7
Price range-high				
139 3/4	123 1/8	130 7/8	129 1/2	175 7/8
Price range-low				
83 1/2	94 1/2	93 3/8	104 1/4	102

7-YEAR PRICE SCORE: 54.40 12-MONTH PRICE SCORE: 64.61

NYSE COMPOSITE INDEX=100

CAPITALIZATION (12/31/91):

	(\$000)	(%)
Long Term Debt	13,231,000	25.4
Deferred Income Tax	1,927,000	3.7
Common & Surplus	37,006,000	70.9
Total	52,164,000	100.0

INSTITUTIONAL SHARES: 262,516,211

INSTITUTIONAL HOLDERS: 1,197

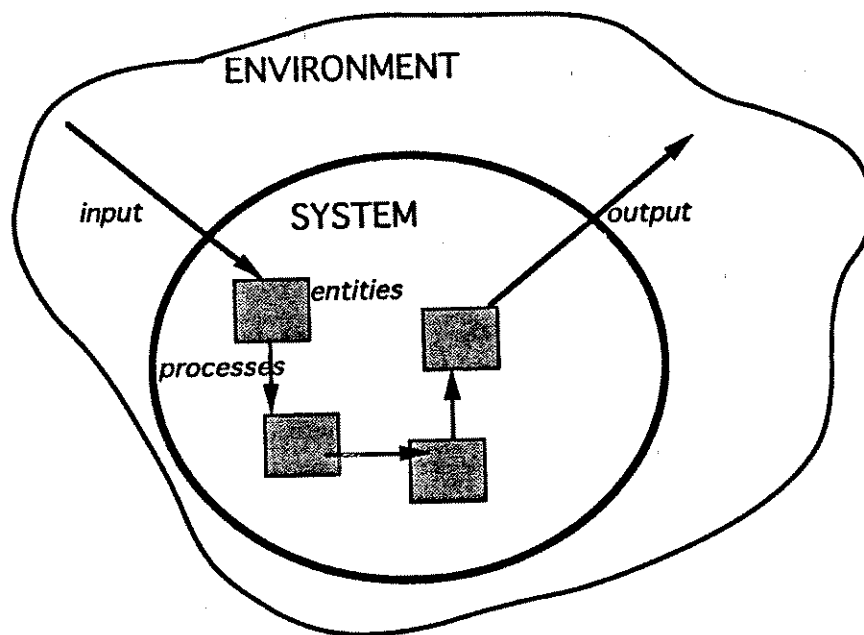
NUMBER OF COMMON STOCKHOLDERS: 772,047

Examples such as these could be almost endless. They are included here to give a feeling for the depth and breadth of information available from online databases. They should help convince you that computer resources and the retrieval skills they require are a vital resource, both for personal information gathering and for the provision of an effective search service to others.

How Does Information Retrieval Work?

Any formal search for information involves some interaction with an information retrieval (IR) system. The word *system* is used to describe a wide variety of phenomena that we encounter in our daily lives—the educational system, the political system—but may best be seen as some set of components that interact to provide a desired result (see fig. 1.13).

Fig. 1.13. A system.



These components consist of a group of interlinked entities (e.g., organizations, people, documents) that participate in a group of interlinked processes (e.g., transmitting, updating, searching). Outside the system, separated from it by a boundary but influencing its operation, lies the system environment. The dynamic nature of the system and the relationships between the elements in it are represented by the information flowing through the system, hence the term *IR cycle*. The transfer of information across the boundary between the system and its environment is known as *input* and *output*. The nature of these inputs, processes, and outputs is governed by the objectives that the system is aiming to fulfill and the external environment within which it operates.

Thus, the typical elements of a document-based IR system consist of inputs and outputs, the matching mechanism, and a series of activities, including:

- the selection of documents;
- the conceptual analysis of documents;
- the organization of *document representations*;
- the storage of documents;
- the conceptual analysis of queries;
- the matching of documents and queries; and
- the delivery of documents.

Let us elaborate on these different elements. The inputs to the system are:

- new documents selected on the basis of user needs;
- ad hoc queries posed by the system users; and
- the indexing language used by both the indexer and the searcher.

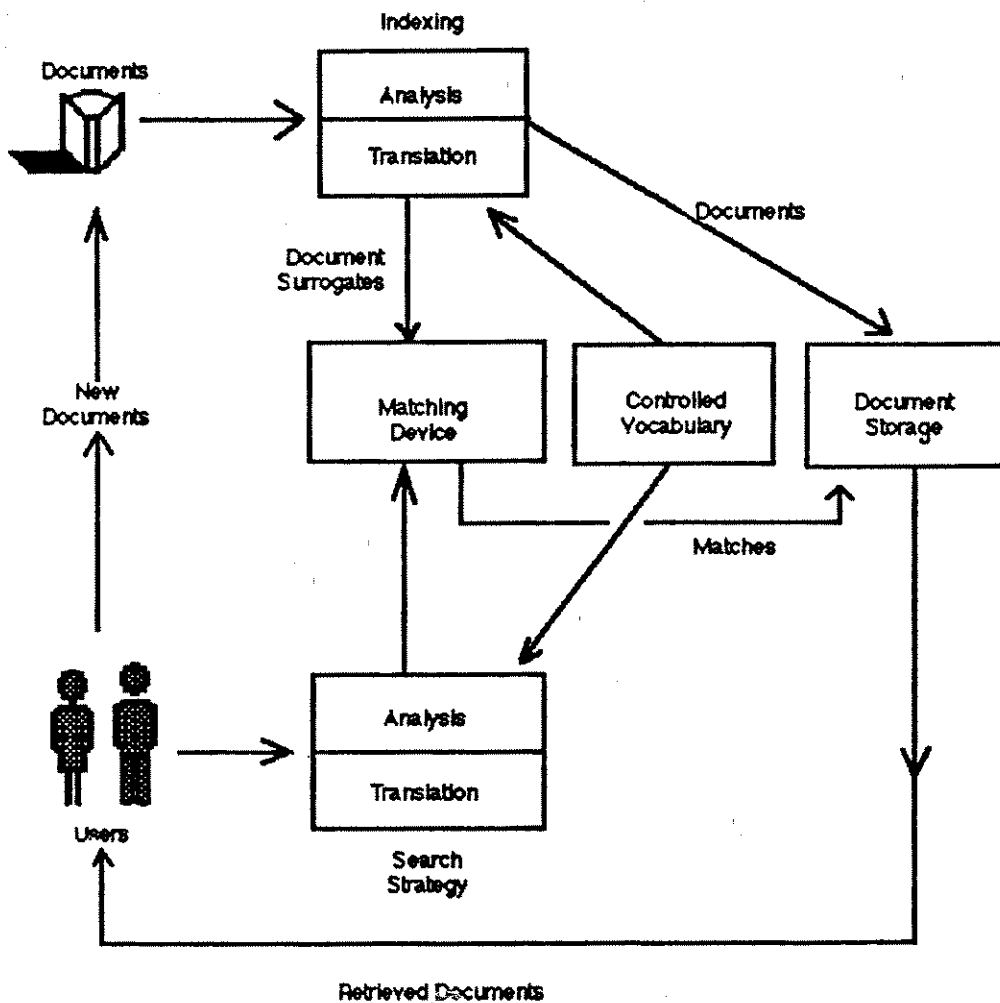
The outputs are:

- documents retrieved in response to queries; and
- factual queries answered.

The relationships among the elements of such an IR system are illustrated in simplified form in figure 1.14 (see p. 15).

And, of course, the cycle is complete when the users who get the documents use them to assist in creating new works, which make their way into the system, and on and on and on. This "cycle of information" takes place not only in the IR system, but everywhere—and is an important underlying phenomenon in library information work. —JWJ

The structure in figure 1.14 is generalized in that it makes no assumptions about how any activity will be carried out, and is equally applicable to either manual or computerized systems. In practice a range of other activities are necessary to link the system inputs and outputs. For example, once the documents have been acquired, they need to be organized in some way so they can be identified and located in response to a search request. This involves cataloging (i.e., description), subject indexing, and (sometimes) abstracting. This indexing consists of two separate stages: the determination of subject content (conceptual analysis) and the translation of that concept into the controlled vocabulary of the system (i.e., the assignment of controlled subject terms). Such a system vocabulary may consist of a list of subject headings, a subject thesaurus, or a classification scheme, and is used to represent the subject matter of the documents for purposes of searching. Most computer systems also make use of *keywords*, or uncontrolled natural language terms that occur in the documents themselves; these will also be searchable. In these systems, both terms from the controlled vocabulary and the keywords may thus be considered subject terms that are searched as representing the documents themselves.

Fig. 1.14. The IR cycle (adapted from Lancaster²).

Indexing is a simple term that sometimes causes confusion. The reason for this is probably that it is used in two rather different ways. We use it to talk of assigning index terms from a controlled vocabulary, but we know that one can also search index terms—keywords—from natural language. Further confusion may be added by the fact that classification numbers are also often considered to be a kind of indexing (subject description). Perhaps we can best summarize this by saying that indexing involves any type of subject specification assigned to documents in order to assist with their retrieval. —GW

Other pieces of information about documents (author's name, date of publication, language, etc.) may also be incorporated into the document representation and thus become available for searching. These document representations (or surrogates) are in fact little summaries of the basic characteristics of the documents being input to the system. They are, of course, much easier to search than the documents themselves because of their condensed nature and their ability to be searched in a variety of ways.

Once the indexing process has been completed, the documents are put into storage, and the document representations are entered into the matching mechanism. This file of document surrogates may be as simple as a card file or a printed index, but for the purposes of online retrieval it is a machine-readable file (a *database*) stored on a computer system. The aim is to make the file conveniently searchable on any search keys deemed to be potentially useful as access points to the document. In the case of the library catalog, the available search keys have traditionally been limited to the author, the title, and a single subject heading. Thus, one of the major advantages of computerized IR systems is their ability to permit searching on a much wider range of document keys—by date, by keyword, by language, and so on.

The input from the other end of the system, the user query, is treated in a fashion as similar as possible to that used for inputting the documents. Queries are analyzed for conceptual content, which is then translated into the vocabulary of the system. This translated version of the query (often with natural-language keywords or other desired document attributes added) becomes the *search strategy*, to be used for matching against the document surrogates in the database.

The output from a document-based IR system will consist of a set of records that the computer has found to match the search strategy. Such records are deemed by the system to be relevant to the user's information need. The success of the retrieval process is usually judged by this attribute of *relevance*, which is generally accepted to be qualitative in nature and uncertain in definition. (For more on relevance, see chapter 12.) The search process, which may be iterative, is finally completed when the user is satisfied with the results of the search, or when it becomes clear that there are no more relevant documents to be found.

A number of researchers have taken a new look at relevance over the past several years. We'll deal with it in more detail in the chapter on evaluation, but for the moment you should know that it appears to be a subtle but measurable phenomenon, dependent on the individual person and situation, and more complicated than one might expect. —JWJ

Information Seeking

It is widely believed that user requests for information usually fall into one of two broad categories:

1. The need to locate and obtain a particular document for which the author or title is known, usually called a *known item search*.
2. The need to locate material dealing with a particular subject or to answer a particular question, known as a *subject search*.

Most information services need to be able to provide answers to both of these types of query. The bibliographic networks (such as OCLC and RLIN) are excellent examples of known item search systems because they are normally used to provide cataloging data for items that are already in hand.

Unfortunately we will be coming across acronyms like this all the time. They are a professional pitfall that you will need to negotiate. OCLC and RLIN are large-scale (many millions of records) bibliographic databases used by many libraries to assist in cataloging and inter-library loan functions. They are known as bibliographic utilities. The OCLC database is produced by the Online Computer Library Center in Ohio, and RLIN is a product of the Research Libraries Group. —GW

The use of these bibliographic utilities is fairly straightforward because they can be searched on accurate and relatively stable elements, such as the ISBN (International Standard Book Number) or an author's name. Subject searching is far more difficult because one is in effect searching for what is not known and which may not even exist. In such cases users are looking for information that will fill a gap in their personal conceptual frameworks, which are probably unique.

A couple of different research approaches have been tried regarding users and gaps in their knowledge states. Belkin, Oddy, and Brooks,³ in a project now known as ASK, studied the anomalous states of knowledge exhibited by users when approaching information systems. They contend that because users have perceived a gap in their knowledge state, they will be "unable to state precisely what is needed to resolve that anomaly." Instead of being asked what information they wanted, users were asked to give a problem statement describing their current knowledge state and the researchers attempted to derive a structural representation of that state to assist in retrieval.

Dervin⁴ has developed a similar sense-making model, which states that people are constantly trying to make sense of their situation as they move through time and space. When they find they are unable to continue moving due to a gap in their sense of the world, they express that gap as an information need, which must be filled for the gap to be resolved and the person to continue on. Both of these theories have been very useful in thinking about and designing information systems. —JWJ

The problems involved in trying to answer this type of subject search arise not only from the difficulty in defining the question but also from the intangible nature of information itself. Information is difficult to define accurately; it is perceived differently by different individuals, and their perceptions are likely to change over time. Although information can reasonably be regarded as anything that helps to answer an information need and may be presented in any way—oral, written, graphic—or in any format—print, microform, computer data, etc.—the mere provision of information-bearing documents does not necessarily mean that information has been effectively transferred.

The growth of funded research and the escalation of publication rates in recent years have posed enormous problems for those whose task it is to acquire and organize the files of published information that have resulted. The organization and processing of this

information is far from simple, and the increased amounts of it have merely served to exacerbate a long-standing conceptual problem. As early as 1960 Maron and Kuhns⁵ highlighted the difficulties involved in identifying the subject content of either documents or queries, because such decisions are not only restricted by the nature of the controlled vocabulary in use but are also bound to contain an element of subjectivity. This was confirmed by Lancaster's early research on the MEDLARS search system,⁶ which suggested that search vocabulary and human errors were the major causes of retrieval failure. The move to computerized systems has undoubtedly speeded up the mechanical parts of IR but has done little to help with the conceptual problems involving subject description.

In fact, the term *information retrieval* itself is something of a misnomer, because what is retrieved by most IR systems is usually either a set of documents or citations to documents that are believed to contain the required information. For example, a library catalog can be searched by subject terms to retrieve records related to the subject required. But the documents identified must then be tracked to the shelves and further searched for particular items of information. The catalog entries are used for the purpose of identifying potentially useful items because they are easier to search than the shelves themselves, but they do not necessarily provide the requested information.

Notice that in this situation we are matching the terms that represent our required subject with the terms that represent the content of the documents. Thus the IR system (in this case the library catalog) is a *matching device* for comparing individual words or phrases between documents and queries. Crucial to the success of this type of retrieval is the vocabulary used to index the documents and to search the document surrogates. It is clear that the same vocabulary has to be used by both the indexer and the searcher, or no matches will be found. This standardization is the role of a *controlled vocabulary*. For example, if we are searching a file for the term "motor cars" when the indexer has entered documents on this subject under the heading "automobiles," we will retrieve nothing, although relevant material is available in the collection. Notice also that although an IR system does not usually retrieve information, information retrieval (IR) is the term commonly used to describe many types of literature searching. Many of the newer systems, however, do provide direct access to the information itself.

As we shall see in chapter 2, the growing problems faced by the traditional search systems (the card catalogs and printed indexes) as they attempted to cope with escalating publication rates led to the early experiments with automated and semiautomated IR systems during the 1950s.

Notes

1. Peter Briscoe, et al. (March 1986), "Ashurbanipal's Enduring Archetype: Thoughts on the Library's Role in the Future," *College & Research Libraries* 47(2): 121-26.
2. F. Wilfred Lancaster (1979), *Information Retrieval Systems: Characteristics, Testing and Evaluation*, 2d ed. (New York: Wiley), 8.
3. N. J. Belkin, R. N. Oddy, and H. M. Brooks (1982), "ASK for Information Retrieval: Part I. Background and Theory," *Journal of Documentation* 38(2): 61-71; and N. J. Belkin, R. N. Oddy, and H. M. Brooks (1982), "ASK for Information Retrieval: Part II. Results of a Design Study," *Journal of Documentation* 38(3): 145-64.

4. B. Dervin (1983), "An Overview of Sense-Making Research: Concepts, Methods and Results to Date." Paper presented at the International Communication Association Annual Meeting, Dallas, TX; and B. Dervin and P. Dewdney (1986), "Neutral Questioning: A New Approach to the Reference Interview," *RQ* 25: 506-13.
5. M. E. Maron and J. L. Kuhns (1960), "On Relevance, Probabilistic Indexing and Information Retrieval," *Journal of Association of Computing Machinery* 7(3): 216-44.
6. F. W. Lancaster (1968), *Evaluation of the MEDLARS Demand Search Service* (Bethesda, MD: National Library of Medicine): 193.

Additional Reading

Lancaster, F. W. (1979), *Information Retrieval Systems: Characteristics, Testing and Evaluation*. 2d ed. (New York: John Wiley): chapters 1, 4, and 5.