Chapter 9

INFORMATION RETRIEVAL AND DIGITAL LIBRARIES

William R. Hersh

Department of Medical Informatics & Clinical Epidemiology, School of Medicine, Oregon Health and Science University, Portland, OR 97239

Chapter Overview

The field of information retrieval (IR) is generally concerned with the indexing and retrieval of knowledge-based information. Although the name implies the retrieval of any type of information, the field has traditionally focused on retrieval of text-based documents, reflecting the type of information that was initially available by this early application of computer use. However, with the growth of multimedia content, including images, video, and other types of information, IR has broadened considerably. The proliferation of IR systems and on-line content has also changed the notion of libraries, which have traditionally been viewed as buildings or organizations. However, the developments of the Internet and new models for publishing have challenged this notion as well, and new digital libraries have emerged.

Keywords

Information retrieval; digital library; indexing; controlled vocabulary; searching; knowledge-based information

Formatted: Bullets and Numbering

1. OVERVIEW OF FIELDS

IR systems and digital libraries store and disseminate knowledge-based information (Hersh, 2003). What exactly do we mean by "knowledge-based"? Although there are many ways to classify biomedical information, for the purposes of this chapter we broadly divide it into two categories. *Patient-specific information* applies to individual patients. Its purpose is to inform health care providers, administrators, and researchers about the health and disease of a patient. This information typically comprises the patient's medical record. The other category of biomedical information is *knowledge-based information*. This information forms the scientific foundation of biomedicine and is derived and organized from observational and experimental research. In the clinical setting, this information provides clinicians, administrators, and researchers with knowledge that can be applied to individual patients. In the basic science (or really any scientific) setting, knowledge-based information provides the archive of research reports upon which further research builds.

Knowledge-based information is most commonly provided in scientific journals and proceedings but can be published in a wide variety of other forms, including books, clinical practice guidelines, consumer health literature, Web sites, and so forth. Figure 9-1 depicts the "life cycle" of primary literature, which is derived from original research and whose publication is dependent upon the peer review process that insures the methods, results, and interpretation of results meets muster with one's scientific peers. In some fields, such as genomics, there is an increasing push for original data to enter public repositories. In most fields, primary information is summarized in secondary publications, such as review articles and textbooks. Also in most fields, the authors relinquish the copyright of their papers to publishers, although there is increasing resistance to this, as described later in this chapter.

IR systems have usually, although not always, been applied to knowledge-based information, which can be subdivided in other ways. *Primary knowledge-based information* (also called primary literature) is original research that appears in journals, books, reports, and other sources. This type of information reports the initial discovery of health knowledge, usually with either original data or re-analysis of data (e.g., meta-analyses).

Secondary knowledge-based information consists of the writing that reviews, condenses, and/or synthesizes the primary literature. As seen in Figure 9-1, secondary literature emanates from original publications. The most common examples of this type of literature are books, monographs, and review articles in journals and other publications. Secondary literature

also includes opinion-based writing such as editorials and position or policy papers.

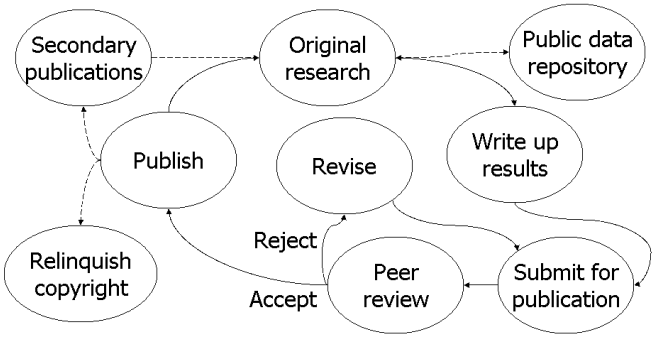

Figure 9-1. The "life cycle" of scientific information.

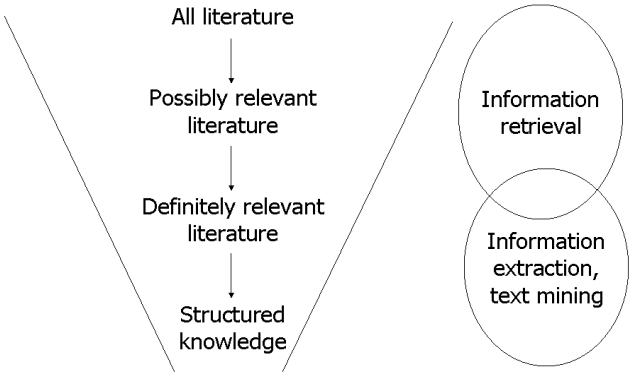

Figure 9-2. Information retrieval and extraction in context.

IR is a distinct process from information extraction (IE), which is covered in many subsequent chapters of this book dealing with vocabularies and ontologies, natural language processing, and text mining. A perspective of the role of IR is provided in Figure 9-2, which shows the flow of extracting knowledge from the scientific literature. IR typically focuses on the initial narrowing of the broad literature, ideally passing off a more focused set of articles for the more intensive processing required for IE and

text mining. A goal for the latter processes is often to create structured knowledge resources that can be accessed by other informatics applications.

Libraries have been the historical place where knowledge-based information has been stored.Libraries actually perform a variety of functions including the following:

- Acquisition and maintenance of collections
- Cataloging and classification of items in collections to make them more accessible to users
- Serving as a place where individuals can go to seek information with assistance, including information on computers
- Providing work or studying space (particularly in universities)

Digital libraries provide some of these same services, but they tend to be more focused on content, particularly in digital form, as opposed to a location, although most physical libraries offer increasing amounts of digital library services (Humphreys, 2000).

2. INFORMATION RETRIEVAL

Now that we have had a general overview of knowledge-based biomedical information, we can look in further detail at IR systems. A model for the IR system and the user interacting with it is shown in Figure 9-3 (Hersh, 2003). The ultimate goal of a user of an IR system is to access content, which may be in the form of a digital library. In order for that content to be accessible, it must be described with metadata. The major intellectual processes of IR are *indexing* and *retrieval*. In the remainder of this section, we will discuss content, indexing, and retrieval, followed by an overview of how IR systems are evaluated.

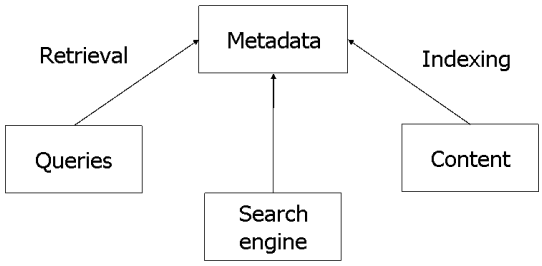

Figure 9-3. A graphic representation of the information retrieval process (Hersh, 2003).

2.1 Content

The ultimate goal of IR systems and digital libraries is to deliver information to users for specific tasks. It is useful to classify the different types of knowledge-based information to better understand the issues in its indexing and retrieval. In this section, we classify content into bibliographic, full-text, database/collection, and aggregated categories and provide an overview of each.

2.1.1 Bibliographic

The first category consists of *bibliographic content*. It includes what was for decades the mainstay of IR systems: literature reference databases. Also called *bibliographic databases*, this content consists of citations or pointers to the medical literature (i.e., journal articles). The best-known and most widely used biomedical bibliographic database is MEDLINE, which is produced by the National Library of Medicine (NLM) and contains bibliographic references to the biomedical articles, editorials, and letters to the editors in approximately 4,500 scientific journals. At present, about 500,000 references are added to MEDLINE yearly. It now contains over 12 million references.

The current MEDLINE record contains up to 49 fields. Probably the most commonly used fields are the title, abstract, and indexing terms. But other fields contain specific information that may be of great importance to smaller audiences. For example, a genomics researcher might be highly interested in the Supplementary Information (SI) field to link to genomic databases. Likewise, the Publication Type (PT) field can be of help to clinicians, designating whether an article is a practice guideline or randomized controlled trial. The NLM also partitions MEDLINE into subsets for users wishing to search on a focused portion of the database, such as *AIDS* or *Complementary and Alternative Medicine*.

MEDLINE is only one of many databases produced by the NLM (Anonymous, 2000c). Other more specialized databases are also available, covering topics from AIDS to space medicine and toxicology. There are a variety of non-NLM bibliographic databases that tend to be more focused on subjects or resource types. The major non-NLM database for the nursing field is CINAHL (Cumulative Index to Nursing and Allied Health Literature, CINAHL Information Systems, www.cinahl.com), which covers nursing and allied health literature, including physical therapy, occupational therapy, laboratory technology, health education, physician assistants, and medical records. Another database is *Excerpta Medica* (Elsevier Science Publishers,

www.excerptamedica.com). EMBASE, the electronic version of *Excerpta Medica*, contains over 8 million records dating back to 1974.

A second, more modern type of bibliographic content is the Web catalog. There are increasing numbers of such catalogs, which consist of Web pages containing mainly links to other Web pages and sites. It should be noted that there is a blurry distinction between Web catalogs and aggregations (the fourth category in this classification). In general, the former contain only links to other pages and sites, while the latter include actual content that is highly integrated with other resources. Some well-known Web catalogs include:

- HealthWeb (healthweb.org)—topics maintained by a consortium of 12 midwestern universities (Redman, Kelly et al., 1997)
- HealthFinder (healthfinder.gov)—consumer-oriented health information maintained by the Office of Disease Prevention and Health Promotion of the U.S. Department of Health and Human Services

There are a number of large general Web catalogs that are not limited to health topics. Two examples are Yahoo (www.yahoo.com) and Open Directory (dmoz.org), both of which have significant health components.

The final type of bibliographic content is the specialized registry. This resource is very close to a literature reference database except that it indexes more diverse content than scientific literature. One specialized registry of great importance for clinicians is the *National Guidelines Clearinghouse* (NGC, www.guideline.gov). Produced by the Agency for Healthcare Research and Quality (AHRQ), it is a bibliographic database with exhaustive information about clinical practice guidelines.

2.1.2 Full-text

The second type of content is *full-text content*. A large component of this content consists of the online versions of books and periodicals. A wide variety of the traditional paper-based biomedical literature, from textbooks to journals, is now available electronically. The electronic versions may be enhanced by measures ranging from the provision of supplemental data in a journal article to linkages and multimedia content in a textbook. The final component of this category is the Web site. Admittedly the diversity of information on Web sites is enormous, and sites may include every other type of content described in this chapter. However, in the context of this category, "Web site" refers to a localized collection (that may be large) of static and dynamic pages at a discrete Web location.

Most biomedical journals are now published in electronic form. Electronic publication not only allows easier access, but additional features

not possible in print versions. For example, journal Web sites can provide additional data with additional figures and tables, results, images, and even raw data. A journal Web site also allows more dialogue about articles than could be published in a Letters to the Editor section of a print journal. Electronic publication also allows true bibliographic linkages, both to other full-text articles and to the MEDLINE record. The Web also allows linkage directly from bibliographic databases to full text. In fact, the MEDLINE database now has a field for the Web address of the full-text paper.

Several hundred biomedical journals use Highwire Press (www.highwire.org) to provide on-line access to their content. The Highwire system provides a retrieval interface that searches over the complete online contents for a given journal. Users can search for authors, words limited to the title and abstract, words in the entire article, and within a date range. The interface also allows searching by citation by entering volume number and page, as well as searching over the entire collection of journals that use Highwire. Users can also browse through specific issues as well as collected resources.

The most common full-text secondary literature source is the traditional textbook, an increasing number of which are available in electronic form. A common approach with textbooks is to bundle multiple books, sometimes with linkages across them. An early bundler of textbooks was *Stat!-Ref* (Teton Data Systems, www.statref.com), which like many began as a CD-ROM product and then moved to the Web. An early product that implemented linking across books was Harrison's Online (McGraw-Hill, www.harrisonsonline.com), which contains the full text of *Harrison's Principles of Internal Medicine* and the drug reference *Gold Standard Pharmacology*. Another textbook collection of growing stature is the *NCBI Bookshelf*, which contains many volumes on biomedical research topics (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books). Some books, such as *On-Line Mendelian Inheritance in Man* (OMIM, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) have ceased publishing paper copies.

Electronic textbooks offer additional features beyond text from the print version. While many print textbooks do feature high-quality images, electronic versions offer the ability to have more pictures and illustrations. They also have the ability to use sound and video, although few do at this time. As with full-text journals, electronic textbooks can link to other resources, including journal references and the full articles. Many Web-based textbook sites also provide access to continuing education self-assessment questions and medical news. In addition, electronic textbooks let authors and publishers provide more frequent updates of the information than is allowed by the usual cycle of print editions, where new versions

come out only every 2 to 5 years.

As noted above, Web sites are another form of full-text information. Probably the most effective user of the Web to provide health information is the U.S. government. The bibliographic databases of the NLM and AHRQ have already been described. These and other agencies, such as the National Cancer Institute (NCI) and Centers for Disease Control and Prevention (CDC) have also been innovative in providing comprehensive full-text information for healthcare providers and consumers as well. Some of these will be described later as aggregations, since they provide many different types of resources. In addition, a large number of private consumer health Web sites have emerged in recent years. Of course they include more than just collections of text, but also interaction with experts, online stores, and catalogs of links to other sites. There are also Web sites that provide information geared toward healthcare providers as well as scientists.

2.1.3 Databases/Collections

The third category consists of *databases and other specific collections* of content. These resources are usually not stored as freestanding Web pages but instead are often housed in database management systems. This content can be further subcategorized into discrete information types:

- Image databases—collections of images from radiology, pathology, and other areas
- Genomics databases—information from gene sequencing, protein characterization, and other genomics research
- Citation databases—bibliographic linkages of scientific literature
- Evidence-based medicine (EBM) databases—highly structured collections of clinical evidence
- Other databases—miscellaneous other collections

A great number of image databases are available on the Web, particularly those from the "visual" medical specialties, such as radiology, pathology, and dermatology. One collection of note is the Visible Human Project of the NLM, which consists of three-dimensional representations of normal male and female bodies (Spitzer, Ackerman et al., 1996). This resource is built from cross-sectional slices of cadavers, with sections of 1 mm in the male and 0.3 mm in the female. Also available from each cadaver are transverse computerized tomography (CT) and magnetic resonance (MR) images. In addition to the images themselves, a variety of searching and browsing interfaces have been created which can be accessed via the project Web site

(http://www.nlm.nih.gov/research/visible/visible_human.html).

Many genomics databases are available across the Web. Some of these are text-based, but even those that are not (such as sequence or structure databases) often contain textual annotations of their data. A key attribute of these databases is their linkage across the Web, such that a record in one database about a gene may have a link to a sequence database with its nucleotide or amino acid sequence, a structure database with the structure of its protein product, or a literature database with papers describing experiments describing the gene. The first issue each year of the journal *Nucleic Acids Research* catalogs and describes these databases (Baxevanis, 2003). At the center of this network of databases are those produced by the National Center for Biotechnology Information (NCBI). All of NCBI's databases are linked among themselves, along with PubMed and OMIM, and are searchable via the Entrez system (http://www.ncbi.nlm.nih.gov/Entrez).

Citation databases provide linkages to articles that cite others across the scientific literature. The best-known citation databases are the *Science Citation Index* (SCI, ISI Thompson) and *Social Science Citation Index* (SSCI, ISI Thompson). A recent development is the *Web of Science*, a Web-based interface to these databases. Another system for citation indexing is the *Research Index* (formerly called *CiteSeer*, citeseer.nj.nec.com) (Lawrence, Giles et al., 1999). This index uses a process called *autonomous citation indexing* that adds citations into its database by automatically processing papers from the Web. It also attempts to identify the context of citations, showing words similar across citations such that the commonality of citing papers can be observed.

EBM databases are devoted to providing synopses of evidence-based information in forms easily accessible by clinicians. Some examples of these databases include the *Cochrane Database of Systematic Reviews*, one of the original collections of systematic reviews (www.cochrane.org), and *Clinical Evidence*, an "evidence formulary" (www.clinicalevidence.com).

There are a variety of other databases/collections that do not fit into the above categories, such as the *ClinicalTrials.gov* database that contains details of ongoing clinical trials sponsored by the National Institutes of Health.

2.1.4 Aggregations

The final category consists of *aggregations* of content from the first three categories. The distinction between this category and some of the highly linked types of content described above is admittedly blurry, but aggregations typically have a wide variety of different types of information

serving diverse needs of their users. Aggregated content has been developed for all types of users from consumers to clinicians to scientists.

Probably the largest aggregated consumer information resource is MEDLINEplus (medlineplus.gov) from the NLM (Miller, Lacroix et al., 2000). MEDLINEplus includes all of the types of content previously described, aggregated for easy access to a given topic. At the top level, MEDLINEplus contains health topics, drug information, medical dictionaries, directories, and other resources. MEDLINEplus currently contains over 400 health topics. The selection of topics is based on analysis of those used by consumers to search for health information on the NLM Web site (Miller, Lacroix et al., 2000). Each topic contains links to health information from the NIH and other sources deemed credible by its selectors. There are also links to current health news (updated daily), a medical encyclopedia, drug references, and directories, along with a preformed PubMed search, related to the topic.

Aggregations of content have also been developed for clinicians. *Merck Medicus* (www.merckmedicus.com) was developed by the well-known publisher and pharmaceutical house, is available to all licensed US physicians, and includes such well-known resources as *Harrison's Online*, *MDConsult* (www.mdconsult.com), and *Dxplain* (http://www.lcs.mgh.harvard.edu/dxplain.htm).

There are many aggregations of content for biomedical researchers as well. Probably the best known among these are the *model organism databases* (Perkel, 2003). These databases bring together bibliographic databases, full text, and databases of sequences, structure, and function for organisms whose genomic data has been highly characterized, such as the mouse (Bult, Blake et al., 2004) and *Saccharomyces* yeast (Bahls, Weitzman et al., 2003). Another well-known aggregation of genomics information is the SOURCE (source.stanford.edu) database, which aggregates information from many other sources about individuals genes in species (Diehn, Sherlock et al., 2003).

2.2 Indexing

Most modern commercial IR systems index their content in two ways. In *manual indexing*, human indexers, usually using standardized terminology, assign indexing terms and attributes to documents, often following a specific protocol. Manual indexing is typically done using *controlled vocabularies*, which consist of the set of allowable terms and relationships between them. In *automated indexing*, on the other hand, computers make the indexing assignments, usually limited to breaking out each word in the document (or part of the document) as an indexing term.

Manual indexing is used most commonly with bibliographic databases. In this age of proliferating electronic content, such as online textbooks, practice guidelines, and multimedia collections, manual indexing has become either too expensive or outright unfeasible for the quantity and diversity of material now available. Thus there are increasing numbers of databases that are indexed only by automated means.

2.2.1 Controlled Vocabularies

Before discussing specific vocabularies it is useful to define some terms, since different writers attach different definitions to the various components of thesauri. A *concept* is an idea or object that occurs in the world, such as the condition under which human blood pressure is elevated. A *term* is the actual string of one or more words that represent a concept, such as *Hypertension* or *High Blood Pressure*. One of these string forms is the *preferred* or *canonical* form, such as *Hypertension* in the present example. When one or more terms can represent a concept, the different terms are called *synonyms*.

A controlled vocabulary usually contains a list of terms that are the canonical representations of the concepts. They are also called *thesauri* and contain relationships between terms, which typically fall into three categories:

- Hierarchical—terms that are broader or narrower. The hierarchical organization not only provides an overview of the structure of a thesaurus but also can be used to enhance searching.
- Synonymous—terms that are synonyms, allowing the indexer or searcher to express a concept in different words.
- Related—terms that are not synonymous or hierarchical but are somehow otherwise related. These usually remind the searcher of different but related terms that may enhance a search.

The Medical Subject Headings (MeSH) vocabulary is used to manually index most of the databases produced by the NLM (Coletti and Bleich, 2001). The latest version contains over 21,000 *subject headings* (the word MeSH uses to denote the canonical representation of its concepts). It also contains over 100,000 supplementary concept records in a separate chemical thesaurus. In addition, MeSH contains the three types of relationships described in the previous paragraph:

- Hierarchical—MeSH is organized hierarchically into 15 trees, such as *Diseases*, *Organisms*, and *Chemicals and Drugs*.

- Synonymous—MeSH contains a vast number of entry terms, which are synonyms of the headings.
- Related—terms that may be useful for searchers to add to their searches when appropriate are suggested for many headings.

The MeSH vocabulary files, their associated data, and their supporting documentation are available on the NLM's MeSH Web site (www.nlm.nih.gov/mesh/). There is also a browser that facilitates exploration of the vocabulary (www.nlm.nih.gov/mesh/MBrowser.html).

There are features of MeSH designed to assist indexers in making documents more retrievable (Anonymous, 2000b). One of these is subheadings, which are qualifiers of subject headings that narrow the focus of a term. Under *Hypertension*, for example, the focus of an article may be on the diagnosis, epidemiology, or treatment of the condition. Another feature of MeSH that helps retrieval is check tags. These are MeSH terms that represent certain facets of medical studies, such as age, gender, human or nonhuman, and type of grant support. Related to check tags are the geographical locations in the Z tree. Indexers must also include these, like check tags, since the location of a study (e.g., *Oregon*) must be indicated. Another feature gaining increasing importance for EBM and other purposes is the publication type, which describes the type of publication or the type of study. A searcher who wants a review of a topic may choose the publication type *Review* or *Review Literature*. Or, to find studies that provide the best evidence for a therapy, the publication type *Meta-Analysis*, *Randomized Controlled Trial*, or *Controlled Clinical Trial* would be used.

MeSH is not the only thesaurus used for indexing biomedical documents. A number of other thesauri are used to index non-NLM databases. CINAHL, for example, uses the *CINAHL Subject Headings*, which are based on MeSH but have additional domain-specific terms added (Brenner and McKinin, 1989). EMBASE has a vocabulary called EMTREE, which has many features similar to those of MeSH (www.elsevier.nl/homepage/sah/spd/site/locate_embase.html).

2.2.2 Manual Indexing

Manual indexing of bibliographic content is the most common and developed use of such indexing. Bibliographic manual indexing is usually done by means of a controlled vocabulary of terms and attributes. Most databases utilizing human indexing usually have a detailed protocol for assignment of indexing terms from the thesaurus. The MEDLINE database is no exception. The principles of MEDLINE indexing were laid out in the two-volume MEDLARS Indexing Manual (Charen, 1976; Charen, 1983). Subsequent modifications have occurred with changes to MEDLINE, other

databases, and MeSH over the years (Anonymous, 2000a). The major concepts of the article, usually from two to five headings, are designed as central concept headings, and designated in the MEDLINE record by an asterisk. The indexer is also required to assign appropriate subheadings. Finally, the indexer must also assign check tags, geographical locations, and publication types.

Few full-text resources are manually indexed. One type of indexing that commonly takes place with full-text resources, especially in the print world, is that performed for the index at the back of the book. However, this information is rarely used in IR systems; instead, most online textbooks rely on automated indexing (see below). One exception to this is MDConsult, which uses back-of-book indexes to point to specific sections in its online books.

Manual indexing of Web content is challenging. With several billion pages of content, manual indexing of more than a fraction of it is not feasible. On the other hand, the lack of a coherent index makes searching much more difficult, especially when specific resource types are being sought. A simple form of manual indexing of the Web takes place in the development of the Web catalogs and aggregations as described above. These catalogs make not only explicit indexing about subjects and other attributes, but also implicit indexing about the quality of a given resource by the decision of whether to include it in the catalog.

Two major approaches to manual indexing have emerged on the Web, which are not mutually incompatible. The first approach, that of applying metadata to Web pages and sites, is exemplified by the Dublin Core Metadata Initiative (DCMI, www.dublincore.org). The second approach, to build directories of content, is further described below.

Table 9-1. Elements of Dublin Core Metadata.

Element	Definition
DC.title	The name given to the resource
DC.creator	The person or organization primarily responsible for creating the intellectual content of the resource
DC.subject	The topic of the resource
DC.description	A textual description of the content of the resource
DC.publisher	The entity responsible for making the resource available in its present form
DC.date	A date associated with the creation or availability of the resource
DC.contributor	A person or organization not specified in a creator element who has made a significant intellectual contribution to the resource but whose contribution is secondary to any person or organization specified in a creator element

continued

Element	Definition
DC.type	The category of the resource
DC.format	The data format of the resource, used to identify the software and possibly hardware that might be needed to display or operate the resource
DC.identifier	A string or number used to uniquely identify the resource
DC.source	Information about a second resource from which the present resource is derived
DC.language	The language of the intellectual content of the resource
DC.relation	An identifier of a second resource and its relationship to the present resource
DC.coverage	The spatial or temporal characteristics of the intellectual content of the resource
DC.rights	A rights management statement, an identifier that links to a rights management statement, or an identifier that links to a service providing information about rights management for the resource

The goal of the DCMI has been to develop a set of standard data elements that creators of Web resources can use to apply metadata to their content (Weibel, 1996). The specification has defined 15 elements, as shown in Table 9-1 (Anonymous, 1999). The DCMI has been anointed a standard by the National Information Standards Organization (NISO) with the designation Z39.85 (Anonymous, 2001a).

While Dublin Core Metadata was originally envisioned to be included in HTML Web pages, it became apparent that many non-HTML resources exist on the Web and that there are reasons to store metadata external to Web pages. For example, authors of Web pages might not be the best people to index pages or other entities might wish to add value by their own indexing of content. An emerging standard for cataloging metadata is the *Resource Description Framework* (RDF) (Miller, 1998). A framework for describing and interchanging metadata, RDF is usually expressed in XML. Increasingly XML is being used to interchange data between databases and has been designated the preferred interchange format in the *Clinical Document Architecture* of the Health Level-7 (HL7, www.hl7.org) standard (Dolin, Alschuler et al., 2001). RDF also forms the basis of what some call the future of the Web as a repository not only of content but also knowledge, which is also referred to as the *Semantic Web* (Lassila, Hendler et al., 2001). Dublin Core Metadata (or any type of metadata) can be represented in RDF (Beckett, Miller et al., 2000).

Another approach to manually indexing content on the Web has been to create directories of content. The first major effort to create these was the Yahoo! search engine, which created a subject hierarchy and assigned Web sites to elements within it (www.yahoo.com). When concern began to emerge that the Yahoo! directory was proprietary and not necessarily

representative of the Web community at large (Caruso, 2000), an alternative movement sprung up, the Open Directory Project.

Manual indexing has a number of limitations, the most significant of which is inconsistency. Funk and Reid (Funk and Reid, 1983) evaluated indexing inconsistency in MEDLINE by identifying 760 articles that had been indexed twice by the NLM. The most consistent indexing occurred with check tags and central concept headings, which were only indexed with a consistency of 61 to 75%. The least consistent indexing occurred with subheadings, especially those assigned to non-central concept headings, which had a consistency of less than 35%. Manual indexing also takes time. While it may be feasible with the large resources the NLM has to index MEDLINE, it is probably impossible with the growing amount of content on Web sites and in other full-text resources. Indeed, the NLM has recognized the challenge of continuing to have to index the growing body of biomedical literature and is investigating automated and semi-automated means of doing so (Aronson, Bodenreider et al., 2000).

2.2.3 Automated Indexing

In automated indexing, the work is done by a computer. Although the mechanical running of the automated indexing process lacks cognitive input, considerable intellectual effort may have gone into building the automated indexing system. In this section, we will focus on the automated indexing used in operational IR systems, namely the indexing of documents by the words they contain.

Some may not think of extracting all the words in a document as "indexing," but from the standpoint of an IR system, words are descriptors of documents, just like human-assigned indexing terms. Most retrieval systems actually use a hybrid of human and word indexing, in that the human-assigned indexing terms become part of the document, which can then be searched by using the whole controlled vocabulary term or individual words within it. Indeed, most MEDLINE implementations have always allowed the combination of searching on human indexing terms and on words in the title and abstract of the reference. With the development of full-text resources in the 1980s and 1990s, systems that only used word indexing began to emerge. This trend increased with the advent of the Web.

Word indexing is typically done by taking all consecutive alphanumeric characters between white space, which consists of spaces, punctuation, carriage returns, and other nonalphanumeric characters. Systems must take particular care to apply the same process to documents and the user's queries, especially with characters such as hyphens and apostrophes. Some systems go beyond simple identification of words and attempt to assign

weights to words that represent their importance in the document (Salton, 1991).

Many systems using word indexing employ processes to remove common words or conflate words to common forms. The former consists of filtering to remove stop words, which are common words that always occur with high frequency and are usually of little value in searching. The *stop list*, also called a *negative dictionary*, varies in size from the seven words of the original MEDLARS stop list (*and, an, by, from, of, the, with*) to the 250 to 500 words more typically used. Examples of the latter are the 250-word list of van Rijsbergen (vanRijsbergen, 1979), the 471-word list of Fox (Fox, 1992), and the PubMed stop list (Anonymous, 2001c). Conflation of words to common forms is done via *stemming*, the purpose of which is to ensure words with plurals and common suffixes (e.g., *-ed, -ing, -er, -al*) are always indexed by their stem form (Frakes, 1992). For example, the words *cough*, *coughs*, and *coughing* are all indexed via their stem *cough*. Stop word removal and stemming also reduce the size of indexing files and lead to more efficient query processing.

A commonly used approach for term weighting is TF*IDF weighting, which combines the inverse document frequency (IDF) and term frequency (TF). The IDF is the logarithm of the ratio of the total number of documents to the number of documents in which the term occurs. It is assigned once for each term in the database, and it correlates inversely with the frequency of the term in the entire database. The usual formula used is:

$$IDF(term) = \log \frac{\text{number of documents in database}}{\text{number of documents with term}} + 1 \quad (1)$$

The TF is a measure of the frequency with which a term occurs in a given document and is assigned to each term in each document, with the usual formula:

$$TF(term, document) = \text{frequency of term in document} \quad (2)$$

In TF*IDF weighting, the two terms are combined to form the indexing weight, WEIGHT:

$$WEIGHT(term, document) = TF(term, document) * IDF(term) \quad (3)$$

Another automated indexing approach generating increased interest is the use of link-based methods, fueled no doubt by the success of the Google (www.google.com) search engine. This approach gives weight to pages based on how often they are cited by other pages. The *PageRank* algorithm is mathematically complex, but can be viewed as giving more weight to a Web page based on the number of other pages that link to it (Brin and Page, 1998). Thus, the home page of the NLM or a major medical journal is likely

to have a very high PageRank (and presumed to be more "authoritative"), whereas a more obscure page will have a lower PageRank. A whole industry has evolved around improving the PageRank scores of one's Web sites (Anonymous, 2003).

Similar to manual indexing, word-based automated indexing has a number of limitations, including:

- Synonymy—different words may have the same meaning, such as *high* and *elevated*. This problem may extend to the level of phrases with no words in common, such as the synonyms *hypertension* and *high blood pressure*.
- Polysemy—the same word may have different meanings or senses. For example, the word *lead* can refer to an element or to a part of an electrocardiogram machine.
- Content—words in a document may not reflect its focus. For example, an article describing *hypertension* may make mention in passing to other concepts, such as *congestive heart failure*, that are not the focus of the article.
- Context—words take on meaning based on other words around them. For example, the relatively common words *high*, *blood*, and *pressure*, take on added meaning when occurring together in the phrase *high blood pressure*.
- Morphology—words can have suffixes that do not change the underlying meaning, such as indicators of plurals, various participles, adjectival forms of nouns, and nominalized forms of adjectives.
- Granularity—queries and documents may describe concepts at different levels of a hierarchy. For example, a user might query for *antibiotics* in the treatment of a specific infection, but the documents might describe specific antibiotics themselves, such as *penicillin*.

2.3 Retrieval

There are two broad approaches to retrieval. *Exact-match searching* allows the user precise control over the items retrieved. *Partial-match searching*, on the other hand, recognizes the inexact nature of both indexing and retrieval, and instead attempts to return the user content ranked by how close it comes to the user's query. After general explanations of these approaches, we will describe actual systems that access the different types of biomedical content.

2.3.1 Exact-match

In exact-match searching, the IR system gives the user all documents that exactly match the criteria specified in the search statement(s). Since the Boolean operators AND, OR, and NOT are usually required to create a manageable set of documents, this type of searching is often called *Boolean searching*. Furthermore, since the user typically builds sets of documents that are manipulated with the Boolean operators, this approach is also called *set-based searching*. Most of the early operational IR systems in the 1950s through 1970s used the exact-match approach, even though Salton was developing the partial-match approach in research systems during that time (Salton and Lesk, 1965). In modern times, exact-match searching tends to be associated with retrieval from bibliographic databases, while the partial-match approach tends to be used with full-text searching. A more detailed example of an exact-match searching system, PubMed, is provided below.

Typically the first step in exact-match retrieval is to select terms to build sets. Other attributes, such as the author name, publication type, or gene identifier (in the secondary source identifier field of MEDLINE), may be selected to build sets as well. Once the search term(s) and attribute(s) have been selected, they are combined with the Boolean operators. The Boolean AND operator is typically used to narrow a retrieval set to contain only documents about two or more concepts. The Boolean OR operator is usually used when there is more than one way to express a concept. The Boolean NOT operator is often employed as a subtraction operator that must be applied to another set. Some systems more accurately call this the ANDNOT operator.

Some systems allow terms in searches to be expanded by using the *wild-card character*, which adds all words to the search that begin with the letters up until the wild-card character. This approach is also called *truncation*. Unfortunately there is no standard approach to using wild-card characters, so syntax for them varies from system to system. PubMed, for example, allows a single asterisk at the end of a word to signify a wild-card character. Thus the query word `can*` will lead to the words *cancer* and *Candida*, among others, being added to the search. The AltaVista search engine (www.altavista.com) takes a different approach. The asterisk can be used as a wild-card character within or at the end of a word but only after its first three letters. For example, *col*r* will retrieve documents containing *color*, *colour*, and *colder*.

2.3.2 Partial-match

Although partial-match searching was conceptualized in the 1960s, it did not see widespread use in IR systems until the advent of Web search engines in the 1990s. This is most likely because exact-match searching tends to be preferred by "power users" whereas partial-match searching is preferred by novice searchers, the ranks of whom have increased substantially with the growth and popularity of the Web. Whereas exact-match searching requires an understanding of Boolean operators and (often) the underlying structure of databases (e.g., the many fields in MEDLINE), partial-match searching allows a user to simply enter a few terms and start retrieving documents.

The development of partial-match searching is usually attributed to Salton (Salton, 1991). Although partial-match searching does not exclude the use of nonterm attributes of documents, and for that matter does not even exclude the use of Boolean operators (e.g., see (Salton, Fox et al., 1983)), the most common use of this type of searching is with a query of a small number of words, also known as a *natural language query*. Because Salton's approach was based on vector mathematics, it is also referred to as the *vector-space model* of IR. In the partial-match approach, documents are typically ranked by their closeness of fit to the query. That is, documents containing more query terms will likely be ranked higher, since those with more query terms will in general be more likely to be relevant to the user. As a result this process is called relevance ranking. The entire approach has also been called *lexical-statistical retrieval*.

The most common approach to document ranking in partial-match searching is to give each a score based on the sum of the weights of terms common to the document and query. Terms in documents typically derive their weight from the TF*IDF calculation described above. Terms in queries are typically given a weight of one if the term is present and zero if it is absent. The following formula can then be used to calculate the document weight across all query terms:

$$Documentweight = \sum_{all\ queryterms} Weightof\ term\ in\ query * Weightof\ term\ in\ document \quad (4)$$

This may be thought of as a giant OR of all query terms, with sorting of the matching documents by weight. The usual approach is for the system to then perform the same stop word removal and stemming of the query that was done in the indexing process. (The equivalent stemming operations must be performed on documents and queries so that complementary word stems will match.)

2.4 Evaluation

There has been a great deal of research over the years devoted to evaluation of IR systems. As with many areas of research, there is controversy as to which approaches to evaluation best provide results that can assess their searching and the systems they are using. Many frameworks have been developed to put the results in context. One of these frameworks organizes evaluation around six questions that someone advocating the use of IR systems might ask (Hersh and Hickam, 1998):

- Was the system used?
- For what was the system used?
- Were the users satisfied?
- How well did they use the system?
- What factors were associated with successful or unsuccessful use of the system?
- Did the system have an impact on the user's task?

A simpler means for organizing the results of evaluation, however, groups approaches and studies into those which are system-oriented, i.e., the focus of the evaluation is on the IR system, and those which are user-oriented, i.e., the focus is on the user.

2.4.1 System-oriented

There are many ways to evaluate the performance of IR systems, the most widely used of which are the relevance-based measures of recall and precision. These measures quantify the number of relevant documents retrieved by the user from the database and in his or her search. They make use of the number of relevant documents (Rel), retrieved documents (Ret), and retrieved documents that are also relevant (Retrel). *Recall* is the proportion of relevant documents retrieved from the database:

$$Recall = \frac{Retrel}{Rel} \tag{5}$$

In other words, recall answers the question, For a given search, what fraction of all the relevant documents have been obtained from the database?

One problem with Eq. (5) is that the denominator implies that the total number of relevant documents for a query is known. For all but the smallest of databases, however, it is unlikely, perhaps even impossible, for one to succeed in identifying all relevant documents in a database. Thus most

studies use the measure of *relative recall*, where the denominator is redefined to represent the number of relevant documents identified by multiple searches on the query topic.

Precision is the proportion of relevant documents retrieved in the search:

$$Precision = \frac{Retrel}{Ret} \tag{6}$$

This measure answers the question, For a search, what fraction of the retrieved documents are relevant?

One problem that arises when one is comparing systems that use ranking versus those that do not is that nonranking systems, typically using Boolean searching, tend to retrieve a fixed set of documents and as a result have fixed points of recall and precision. Systems with relevance ranking, on the other hand, have different values of recall and precision depending on the size of the retrieval set the system (or the user) has chosen to show. For this reason, many evaluators of systems featuring relevance ranking will create a recall-precision table (or graph) that identifies precision at various levels of recall. The "standard" approach to this was defined by Salton (Salton, 1983), who pioneered both relevance ranking and this method of evaluating such systems.

To generate a recall-precision table for a single query, one first must determine the intervals of recall that will be used. A typical approach is to use intervals of 0.1 (or 10%), with a total of 11 intervals from a recall of 0.0 to 1.0. The table is built by determining the highest level of overall precision at any point in the output for a given interval of recall. Thus, for the recall interval 0.0, one would use the highest level of precision at which the recall is anywhere greater than or equal to zero and less than 0.1. An approach that has been used more frequently in recent times has been the *mean average precision* (MAP), which is similar to precision at points of recall but does not use fixed recall intervals or interpolation (Voorhees, 1998). Instead, precision is measured at every point at which a relevant document is obtained, and the MAP measure is found by averaging these points for the whole query.

No discussion of IR evaluation can ignore the *Text REtrieval Conference* (TREC, trec.nist.gov) organized by the U.S. National Institute for Standards and Technology (NIST, www.nist.gov) (Voorhees and Harman, 2000). Started in 1992, TREC has provided a testbed for evaluation and a forum for presentation of results. TREC is organized as an annual event at which the tasks are specified and queries and documents are provided to participants. Participating groups submit "runs" of their systems to NIST, which calculates the appropriate performance measure, usually recall and precision. TREC is organized into tracks geared to specific interests. Voorhees

recently grouped the tracks into general IR tasks (Voorhees and Harman, 2001):

- Static text—Ad Hoc
- Streamed text—Routing, Filtering
- Human in the loop—Interactive
- Beyond English (cross-lingual)—Spanish, Chinese, and others
- Beyond text—OCR, Speech, Video
- Web searching—Very Large Corpus, Web
- Answers, not documents—Question-Answering
- Retrieval in a domain—Genomics

TREC has been an initiative for the general IR community and, as such, has mostly newswire, government, and Web (i.e., non-biomedical) content. However, a recent track has been formed using biomedical data, the TREC Genomics Track (http://medir.ohsu.edu/~genomics). The first year of the track featured tasks in both IR and IE (Hersh and Bhupatiraju, 2003). Further iterations of the track will feature more advanced approaches to evaluation of retrieval as well as user studies. Another advantage of this track has been to bring the IR and bioinformatics research communities into more contact.

Relevance-based measures have their limitations. While no one denies that users want systems to retrieve relevant articles, it is not clear that the quantity of relevant documents retrieved is the complete measure of how well a system performs (Swanson, 1988; Harter, 1992). Hersh (Hersh, 1994) has noted that clinical users are unlikely to be concerned about these measures when they simply seek an answer to a clinical question and are able to do so no matter how many other relevant documents they miss (lowering recall) or how many nonrelevant ones they retrieve (lowering precision).

What alternatives to relevance-based measures can be used for determining performance of individual searches? Many advocate that the focus of evaluation put more emphasis on user-oriented studies, particularly those that focus on how well users perform real-world tasks with IR systems. Some of these studies are described in the next section, while a series of biomedically focused user studies by Hersh and colleagues are presented later.

2.4.2 User-oriented

A number of user-oriented evaluations have been performed over the years looking at users of biomedical information. Most of these studies have

focused on clinicians. One of the original studies measuring searching performance in clinical settings was performed by Haynes et al. (Haynes, McKibbon et al., 1990). This study also compared the capabilities of librarian and clinician searchers. In this study, 78 searches were randomly chosen for replication by both a clinician experienced in searching and a medical librarian. During this study, each original ("novice") user had been required to enter a brief statement of information need before entering the search program. This statement was given to the experienced clinician and librarian for searching on MEDLINE. All the retrievals for each search were given to a subject domain expert, blinded with respect to which searcher retrieved which reference. Recall and precision were calculated for each query and averaged. The results (Table 9-2) showed that the experienced clinicians and librarians achieved comparable recall, although the librarians had statistically significantly better precision. The novice clinician searchers had lower recall and precision than either of the other groups. This study also assessed user satisfaction of the novice searchers, who despite their recall and precision results said that they were satisfied with their search outcomes. The investigators did not assess whether the novices obtained enough relevant articles to answer their questions, or whether they would have found additional value with the ones that were missed.

Table 9-2. Recall and precision of MEDLINE searchers.

Users	Recall (%)	Precision (%)
Novice clinicians	27	38
Experienced clinicians	48	49
Medical librarians	49	58

A follow-up study yielded some additional insights about the searchers (McKibbon, Haynes et al., 1990). As was noted, different searchers tended to use different strategies on a given topic. The different approaches replicated a finding known from other searching studies in the past, namely, the lack of overlap across searchers of overall retrieved citations as well as relevant ones. Thus, even though the novice searchers had lower recall, they did obtain a great many relevant citations not retrieved by the two expert searchers. Furthermore, fewer than 4% of all the relevant citations were retrieved by all three searchers. Despite the widely divergent search strategies and retrieval sets, overall recall and precision were quite similar among the three classes of users.

Other user-oriented evaluation has looked at how well users complete tasks with IR systems. Egan et al. (Egan, Remde et al., 1989) evaluated the effectiveness of the Superbook application by assessing how well users could find and apply specific information. Mynatt et al. (Mynatt, Leventhal et al., 1992) used a similar approach in comparing paper and electronic

versions of an online encyclopedia, while Wildemuth et al. (Wildemuth, deBliek et al., 1995) assessed the ability of students to answer testlike questions using a medical curricular database. The TREC Interactive Track has also used this approach. This work showed that some algorithms found effective using system-oriented, relevance-based evaluation measures did not maintain that effectiveness in experiments with real users (Hersh, 2001).

2.5 Research Directions

A steady stream of research continues to look at new approaches to IR, a detailed discussion of which is beyond the scope of this chapter. The NLM sponsors biomedical IR research both internally and externally. Its biggest internal project is the *Indexing Initiative*, which is investigating new approaches to automated and semi-automated indexing, mostly based on tools using the UMLS and natural language processing tools (Aronson, Bodenreider et al., 2000).

Other approaches to research have focused on improving aspects of automated indexing and retrieval. A number of these have been found to improve retrieval performance in the TREC environment, including:

- Improved approaches to term weighting, such as Okapi (Robertson and Walker, 1994), pivoted normalization (Singhal, Buckley et al., 1996), and language modeling (Ponte and Croft, 1998)
- Passage retrieval, where documents are given more weight in the ranking process based on local concentrations of query terms within them (Callan, 1994)
- Query expansion, where new terms from highly ranking documents are added to the query in an automated fashion (Srinivasan, 1996; Xu and Croft, 1996)

Additional work has focused on improving the user interface for the retrieval process by organizing the output better. An example of this is *Dynacat*, a system for consumers which uses UMLS knowledge and MeSH terms to organize search results (Pratt, Hearst et al., 1999). The goal is to present search results with documents clustered into topical groups, such as the treatments for a disease or the tests used to diagnose it. Another approach is to make the search system vocabulary more understandable in context. The *Cat-a-Cone* system provides a means to explore term hierarchies by using *cone trees*, which rotate the primary term of interest to the center of the screen and show conelike expansion of other hierarchically related terms nearby (Hearst and Karadi, 1997).

3. DIGITAL LIBRARIES

Discussion of IR "systems" thus far has focused on the provision of retrieval mechanisms to access online content. Even with the expansive coverage of some IR systems, such as Web search engines, they are often part of a larger collection of services or activities. An alternative perspective, especially when communities and/or proprietary collections are involved, is the digital library. Digital libraries share many characteristics with "brick and mortar" libraries, but also take on some additional challenges. Borgman (Borgman, 1999) notes that libraries of both types elicit different definitions of what they actually are, with researchers tending to view libraries as content collected for specific communities and practitioners alternatively viewing them as institutions or services.

3.1 Access

Probably every Web user is familiar with clicking on a Web link and receiving the error message: *HTTP 404 - File not found*. Digital libraries and commercial publishing ventures need mechanisms to ensure that documents have persistent identifiers so that when the document itself physically moves, it is still obtainable. The original architecture for the Web envisioned by the Internet Engineering Task Force was to have every uniform resource locator (URL), the address entered into a Web browser or used in a Web hyperlink, linked to a uniform resource name (URN) that would be persistent (Sollins and Masinter, 1994). The combination of a URN and URL, a uniform resource identifier (URI), would provide persistent access to digital objects. The resource for resolving URNs and URIs was never implemented on a large scale.

One approach that has begun to see widespread adoption by publishers, especially scientific journal publishers, is the digital object identifier (DOI, www.doi.org) (Paskin, 1999). The DOI has recently been given the status of a standard by the National Information Standards Organization (NISO) with the designation Z39.84. The DOI itself is relatively simple, consisting of a prefix that is assigned by the IDF to the publishing entity and a suffix that is assigned and maintained by the entity. For example, the DOI for articles from the *Journal of the American Medical Informatics Association* have the prefix *10.1197* and the suffix *jamia.M####*, where *####* is a number assigned by the journal editors. Likewise, all publications in the Digital Library of the Association for Computing Machinery (http://www.acm.org/dl) have the prefix *10.1145* and a unique identifier for the suffix (e.g., *345508.345539*) for the paper. Publishers are encouraged to facilitate resolution by encoding the DOI into their URLs in a standard way, e.g., http://doi.acm.org/10.1145/345508.345539.

3.2 Interoperability

As noted throughout this chapter, metadata is a key component for accessing content in IR systems. It takes on additional value in the digital library, where there is desire to allow access to diverse but not necessarily exhaustive resources. One key concern of digital libraries is interoperability (Besser, 2002). That is, how can resources with heterogeneous metadata be accessed? Arms et al. (Arms, Hillmann et al., 2002) note that three levels of agreement must be achieved:

- Technical agreements over formats, protocols, and security procedures
- Content agreement over the data and the semantic interpretation of its metadata
- Organizational agreements over ground rules for access, preservation, payment, authentication, and so forth

One approach to interoperability gaining increasing use is the Open Archives Initiative (OAI, www.openarchive.org) (Lagoze and VandeSompel, 2001). While the OAI effort is rooted in access to scholarly communications, its methods are applicable to a much broader range of content. Its fundamental activity is to promote the "exposure" of archives' metadata such that digital library systems can learn what content is available and how it can be obtained. Each record in the OAI system has an XML-encoded record. The OAI Protocol for Metadata Harvesting (PMH) then allows selective harvesting of the metadata by systems. Such harvesting can be date-based, such as items added or changed after a certain date, or set-based, such as those belonging to a certain topic, journal, or institution. A growing number of biomedical resources have adopted OAI (McKiernan, 2003).

3.3 Preservation

Another concern for digital libraries is the preservation of content, especially with the growing trend towards electronic subscriptions to journals that result in fewer physical copies (electronic or printed) being produced. Also a concern is the longevity of digital materials (Lesk, 1997). Of all media, the longevity is the least for magnetic materials, with the expected lifetime of magnetic tape being 5 to 10 years. Optical storage has somewhat better longevity, with an expected lifetime of 30 to 100 years depending on the specific type. Ironically, paper has a life expectancy well beyond all these digital media. A growing concern is that with the increasing move towards electronic publishing, there are fewer copies of journal material produced using media that have lesser longevity.

As such, there is an imperative to preserve documents of many types, whatever their medium (Tibbo, 2001). For society in general, there is

certainly impetus to preserve historical documents in an unaltered form. And in all of science, certainly biomedicine, there is need to preserve the archive of scientific discoveries, particularly those presenting original experiments and their data. A number of initiatives have been undertaken to insure preservation of digital information. These include the National Digital Information Infrastructure Preservation Program (NDIIPP, www.digitalpreservation.gov) of the US Library of Congress (Friedlander, 2002) and the Digital Preservation Coalition in the United Kingdom (Beagrie, 2002).

4. CASE STUDIES

In this section, we will explore three case studies or examples of IR in further detail. These include a retrieval system, user-oriented evaluation, and issues surrounding electronic publishing.

Formatted: Bullets and Numbering

4.1 PubMed

Probably the best known and most widely used biomedical IR system is PubMed (pubmed.gov) from the NLM. (Unless one considers Google to be a biomedical IR system, for which a tenable case can be made!) PubMed searches MEDLINE and other bibliographic databases from the NLM. Although presenting the user with a simple text box, PubMed does a great deal of processing of the user's input to identify MeSH terms, author names, common phrases, and journal names (Anonymous, 2001c). In this automatic term mapping, the system attempts to map user input, in succession, to MeSH terms, journal names, common phrases, and authors. Remaining text that PubMed cannot map is searched as text words (i.e., words that occur in any of the MEDLINE fields).

PubMed allows the use of wild-card characters. It also allows phrase searching in that two or more words can be enclosed in quotation marks to indicate they must occur adjacent to each other. If the specified phrase is in PubMed's phrase index, then it will be searched as a phrase. Otherwise the individual words will be searched. PubMed allows specification of other indexing attributes via the PubMed "Limits" screen. These include publication types, subsets, age ranges, and publication date ranges.

As in most bibliographic systems, users search PubMed by building search sets and then combining them with Boolean operators to tailor the search. Consider a user searching for studies assessing the reduction of mortality in patients with *congestive heart failure (CHF)* through the use of medications from the *angiotensin-converting (ACE) inhibitors* class of

drugs. A simple approach to such a search might be to combine the terms *ACE Inhibitors* and *CHF* with an AND. The easiest way to do this is to enter the search string *ace inhibitors AND CHF*. (The operator *AND* must be capitalized because PubMed treats the lowercase `and` as a text word, since some MeSH terms, such as *Bites and Strings*, have the word `and` in them.) Figure 9-4 shows the PubMed History screen such a searcher might develop.

PubMed also has a *Clinical Queries* interface, where the subject terms are limited by search statements designed to retrieve the best evidence based on principles of EBM. There are two different approaches. The first uses strategies for retrieving the best evidence for the four major types of clinical question. These strategies arise from research assessing the ability of MEDLINE search statements to identify the best studies for therapy, diagnosis, harm, and prognosis (Haynes, Wilczynski et al., 1994). The second approach to retrieving the best evidence aims to retrieve evidence-based resources including meta-analyses, systematic reviews, and practice guidelines. When the Clinical Queries interface is used, the search statement is processed by the usual automatic term mapping and the resulting output is limited (via AND) with the appropriate statement.

PubMed is actually part of the larger Entrez system at NLM that provides access to the entire range of on-line content (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi). Another interface that searches over the range of NLM content is the NLM Gateway (http://gateway.nlm.nih.gov/gw/Cmd).

4.2 User-oriented Evaluation

Recognizing the limitations of recall and precision for evaluating clinical users of IR systems, Hersh and colleagues have carried out a number of studies assessing the ability of systems to help students and clinicians answer clinical questions. The rationale for these studies is that the usual goal of using an IR system is to find an answer to a question. While the user must obviously find relevant documents to answer that question, the quantity of such documents is less important than whether the question is successfully answered. In fact, recall and precision can be placed among the many factors that may be associated with ability to complete the task successfully.

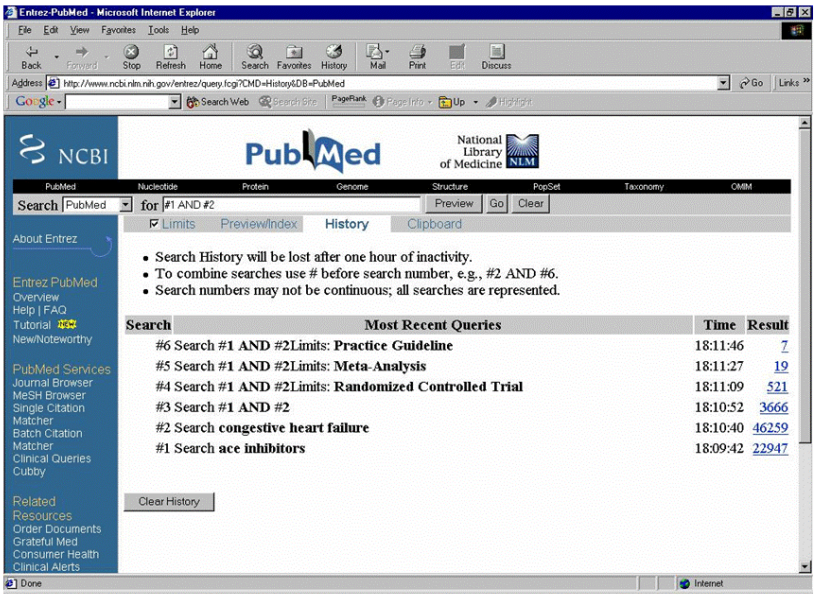

Figure 9-4. PubMed History screen. (Courtesy of NLM.)

The first study by this group using the task-oriented approach compared Boolean versus natural language searching in the textbook *Scientific American Medicine* (Hersh, Elliot et al., 1994). Thirteen medical students were asked to answer 10 short-answer questions and rate their confidence in their answers. The students were then randomized to one or the other interface and asked to search on the five questions for which they had rated confidence the lowest. The study showed that both groups had low correct rates before searching (average 1.7 correct out of 10) but were mostly able to answer the questions with searching (average 4.0 out of 5). There was no difference in ability to answer questions with one interface or the other. Most answers were found on the first search of the textbook. For the questions that were incorrectly answered, the document with the correct answer was actually retrieved by the user two-thirds of the time and viewed more than half the time.

Another study compared Boolean and natural language searching of MEDLINE with two commercial products, CD Plus (now Ovid, www.ovid.com) and Knowledge Finder (Aries Systems, www.ariessystems.com) (Hersh, Pentecost et al., 1996). These systems represented the ends of the spectrum in terms of using Boolean searching on human-indexed thesaurus terms (CD Plus) versus natural language searching

on words in the title, abstract, and indexing terms (Knowledge Finder). Sixteen medical students were recruited and randomized to one of the two systems and given three yes/no clinical questions to answer. The students were able to use each system successfully, answering 37.5% correct before searching and 85.4% correct after searching. There were no significant differences between the systems in time taken, relevant articles retrieved, or user satisfaction. This study demonstrated that both types of system can be used equally well with minimal training.

The most comprehensive study looked at MEDLINE searching by medical and nurse practitioner (NP) students to answer clinical questions. A total of 66 medical and NP students searched five questions each (Hersh, Crabtree et al., 2002). This study used a multiple-choice format for answering questions that also included a judgment about the evidence for the answer. Subjects were asked to choose from one of three answers:

- Yes, with adequate evidence
- Insufficient evidence to answer question
- No, with adequate evidence

Both groups achieved a presearching correctness on questions about equal to chance (32.3% for medical students and 31.7% for NP students). However, medical students improved their correctness with searching (to 51.6%), whereas NP students hardly did at all (to 34.7%).

This study also assessed what factors were associated with successful searching. A number of factors, such as age, gender, computer experience, and time taken to search, were not associated with successful answering of questions. However, successful answering was associated with answering the question correctly before searching, spatial visualization ability (measured by a validated instrument), searching experience, and EBM question type (prognosis questions easiest, harm questions most difficult). An analysis of recall and precision for each question searched demonstrated their complete lack of association with ability to answer these questions.

4.3 Changes in Publishing

Any discussion of IR systems and digital libraries cannot ignore the larger context of the political and economic aspects of publishing. While a complete discussion is beyond the scope of a chapter like this, some of the high points can and should be elucidated, if for no other reason than that they impact access to content for the kinds of innovations and research described in this book.

The Internet and WWW have had profound impact in the publishing of knowledge-based information. The technical impediments to electronic publishing of journals have largely been solved. Most scientific journals are

published electronically in some form already. Journals that do not publish electronically likely could do so easily, since most of the publishing process has already been converted to the electronic mode. A modern Internet connection is sufficient to deliver most of the content of journals. Indeed, a near turnkey solution is already offered through Highwire Press, which has an infrastructure that supports journal publishing from content preparation to searching and archiving.

There is great enthusiasm for electronic availability of journals, as evidenced by the growing number of titles to which libraries provide access. Likewise, since most scientists have the desire for widespread dissemination of their work, they have incentive for their papers to be available on the Web. Indeed, it has been shown, at least in the computer science domain, that papers freely available on the Web have a higher likelihood of being cited by other papers than those which are not (Lawrence, 2001). As citations are important to authors for academic promotion and grant funding, authors have incentive to maximize the accessibility of their published work.

The technical challenges to electronic scholarly publication have been replaced by economic and political ones (Hersh and Rindfleisch, 2000; Anonymous, 2001b). Printing and mailing, tasks no longer needed in electronic publishing, comprised a significant part of the "added value" from publishers of journals. There is, however, still value added by publishers, such as hiring and managing editorial staff to produce the journals and managing the peer review process. Even if publishing companies as they are known were to vanish, there would still be some cost to the production of journals. Thus, while the cost of producing journals electronically is likely to be less, it is not zero, and even if journal content is distributed "free," someone has to pay the production costs.

The economic issue in electronic publishing, then, is who is going to pay for the production of journals. This introduces some political issues as well. One of them centers on the concern that much research is publicly funded through grants from federal agencies such as the National Institutes of Health (NIH) and the National Science Foundation (NSF). In the current system, especially in the biomedical sciences (and to a lesser extent in nonbiomedical sciences), researchers turn over the copyright of their publications to journal publishers. The political concern is that the public funds the research and the universities carry it out, but individuals and libraries then must buy it back from the publishers to whom they willingly cede the copyright (McCook, 2004). This problem is exacerbated by the general decline in funding for libraries that has occurred over the last couple decades (Boyd and Herkovic, 1999; Meek, 2001).

Some have proposed models of scholarly publishing that keep the archive of science freely available. One of these is *open access* publishing, where

authors and their institutions pay the cost of production of manuscripts up front after they are accepted through a peer review process. It has been suggested that this cost could even be included in the budgets of grant proposals submitted for funding agencies. After the paper is published, the manuscript becomes freely available on the Web. The first publisher to take this approach has been Biomed Central (BMC, www.biomedcentral.com). Another highly visible open access approach is the Public Library of Science (PLOS, www.plos.org). Although legislation has been proposed requiring research funded by government agencies to be open access (McLellan, 2003), at least one journal editor has expressed caution that the untested model may not work as well as advertised, especially for the major biomedical journals that devote substantial resources to insuring the quality of high-profile biomedical research (DeAngelis and Musacchio, 2004).

Another model is that of PubMed Central (PMC, pubmedcentral.gov), which provides free access to published literature but allows publishers to maintain copyright as well as optionally keep the papers on their own servers. A lag time of up to 6 months is allowed so that journals can reap the revenue that comes with initial publication. The number of journals submitting their content to PMC has been modest, and there are currently about 100 that contribute to its repository.

5. ACKNOWLEDGEMENTS

The author's research has been generously funded by the National Library of Medicine, Agency for Healthcare Quality and Research, and National Science Foundation over the years. He is particularly grateful to the NLM for its strong leadership in promoting research and education in the field of medical informatics.

REFERENCES

Anonymous. (1999). Dublin Core Metadata Element Set, Version 1.1: Reference Description. Dublin Core Metadata Initiative, http://www.dublincore.org/documents/dces/.

Anonymous. (2000a). Cataloging Practices. National Library of Medicine, http://www.nlm.nih.gov/mesh/catpractices.html.

Anonymous. (2000b). Features of the MeSH Vocabulary. National Library of Medicine, http://www.nlm.nih.gov/mesh/features.html.

Anonymous. (2000c). Organization of National Library of Medicine Bibliographic Databases. National Library of Medicine, http://www.nlm.nih.gov/pubs/techbull/mj00/mj00_buckets.html.

Anonymous. (2001a). The Dublin Core Metadata Element Set. Dublin Core Metadata Initiative, http://www.niso.org/standards/resources/Z39-85.pdf.

Anonymous. (2001b). "The Future of the Electronic Scientific Literature," *Nature*, 413: 1-3.

Anonymous. (2001c). PubMed Help. National Library of Medicine, http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html. Accessed: July 1, 2002.

Anonymous. (2003). The Google Ranking Report. Sedona, AZ, Cyberdifference Corp., http://www.mseo.com/google_ranking_report.html.

Arms, W., Hillmann, D., et al. (2002). "A Spectrum of Interoperability: The Site for Science Prototype for the NSDL," *D-Lib Magazine*, 8, http://www.dlib.org/dlib/january02/arms/01arms.html.

Aronson, A., Bodenreider, O., et al. (2000). "The NLM Indexing Initiative," in *Proceedings of the AMIA 2000 Annual Symposium*, Los Angeles, CA. Hanley & Belfus, 17-21.

Bahls, C., Weitzman, J., et al. (2003). "Biology's Models," *The Scientist*. June 2, 2003. 5, http://www.the-scientist.com/yr2003/jun/feature_030602.html.

Baxevanis, A. (2003). "The Molecular Biology Database Collection: 2003 update," *Nucleic Acids Research*, 31: 1-12.

Beagrie, N. (2002). "An Update on the Digital Preservation Coalition," *D-Lib Magazine*, 8, http://www.dlib.org/dlib/april02/beagrie/04beagrie.html.

Beckett, D., Miller, E., et al. (2000). Using Dublin Core in XML. Dublin Core Metadata Initiative, http://dublincore.org/documents/dcmes-xml/.

Besser, H. (2002). "The Next Stage: Moving from Isolated Digital Collections to Interoperable Digital Libraries," *First Monday*, 7(6), http://www.firstmonday.dk/issues/issue7_6/besser/

Borgman, C. (1999). "What are Digital Libraries? Competing Visions," *Information Processing and Management*, 35: 227-244.

Boyd, S. and Herkovic, A. (1999). Crisis in Scholarly Publishing: Executive Summary. Stanford Academic Council Committee on Libraries, http://www.stanford.edu/~boyd/schol_pub_crisis.html.

Brenner, S. and McKinin, E. (1989). "CINAHL and MEDLINE: A Comparison of Indexing Practices," *Bulletin of the Medical Library Association*, 77: 366-371.

Brin, S. and Page, L. (1998). "The Anatomy of a Large-scale Hypertextual Web Search Engine," *Computer Networks*, 30: 107-117.

Bult, C., Blake, J., et al. (2004). "The Mouse Genome Database (MGD): Integrating Biology with the Genome," *Nucleic Acids Research*, 32: D476-481.

Callan, J. (1994). "Passage Level Evidence in Document Retrieval," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland. Springer-Verlag. 302-310.

Caruso, D. (2000). "Digital Commerce; If the AOL-Time Warner Deal is about Proprietary Content, Where Does that Leave a Noncommercial Directory It Will Own?" *New York Times*. January 17, 2000.

Charen, T. (1976). *MEDLARS Indexing Manual, Part I: Bibliographic Principles and Descriptive Indexing, 1977*. Springfield, VA: National Technical Information Service.

Charen, T. (1983). *MEDLARS Indexing Manual, Part II*. Springfield, VA: National Technical Information Service.

Coletti, M. and Bleich, H. (2001). "Medical Subject Headings Used to Search the Biomedical Literature," *Journal of the American Medical Informatics Association*, 8: 317-323.

DeAngelis, C. and Musacchio, R. (2004). "Access to JAMA," *Journal of the American Medical Association*, 291: 370-371.

Diehn, M., Sherlock, G., et al. (2003). "SOURCE: A Unified Genomic Resource of Functional Annotations, Ontologies, and Gene Expression Data," *Nucleic Acids Research*, 31: 219-223.

Dolin, R., Alschuler, L., et al. (2001). "The HL7 Clinical Document Architecture," *Journal of the American Medical Informatics Association*, 8: 552-569.

Egan, D., Remde, J., et al. (1989). "Formative Design-evaluation of Superbook," *ACM Transactions on Information Systems*, 7: 30-57.

Fox, C. (1992). "Lexical Analysis and Stop Lists," in Frakes, W. and Baeza-Yates, R., eds. *Information Retrieval: Data Structures and Algorithms*, Englewood Cliffs, NJ: Prentice-Hall, pp.102-130,

Frakes, W. (1992). "Stemming Algorithms," in Frankes, W. and Baeza-Yates, R., eds. *Information Retrieval: Data Structures and Algorithms*, Englewood Cliffs, NJ: Prentice-Hall, pp. 131-160.

Friedlander, A. (2002). "The National Digital Information Infrastructure Preservation Program: Expectations, Realities, Choices, and Progress to Date," *D-Lib Magazine*, 8, http://www.dlib.org/dlib/april02/friedlander/04friedlander.html.

Funk, M. and Reid, C. (1983). "Indexing Consistency in MEDLINE," *Bulletin of the Medical Library Association*, 71: 176-183.

Harter, S. (1992). "Psychological Relevance and Information Science," *Journal of the American Society for Information Science*, 43: 602-615.

Haynes, R., McKibbon, K., et al. (1990). "Online Access to MEDLINE in Clinical Settings," *Annals of Internal Medicine*, 112: 78-84.

Haynes, R., Wilczynski, N., et al. (1994). "Developing Optimal Search Strategies for Detecting Clinically Sound Studies in MEDLINE," *Journal of the American Medical Informatics Association*, 1: 447-458.

Hearst, M. and Karadi, C. (1997). "Cat-a-Cone: An Interactive Interface for Specifying Searches and Viewing Retrieval Results Using a Large Category Hierarchy," in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, PA. ACM Press. 246-255.

Hersh, W. (1994). "Relevance and Retrieval Evaluation: Perspectives from Medicine," *Journal of the American Society for Information Science*, 45: 201-206.

Hersh, W. (2001). "Interactivity at the Text Retrieval Conference (TREC)," *Information Processing and Management*, 37: 365-366.

Hersh, W. (2003). *Information Retrieval: A Health and Biomedical Perspective*. Second Edition. New York: Springer-Verlag, http://www.irbook.org.

Hersh, W. and Bhupatiraju, R. (2003). "TREC Genomics track overview," in *The Twelfth Text Retrieval Conference: TREC 2003*, Gaithersburg, MD. National Institute of Standards & Technology, http://trec.nist.gov/pubs/trec12/papers/GENOMICS.OVERVIEW3.pdf.

Hersh, W., Crabtree, M., et al. (2002). "Factors Associated with Success for Searching MEDLINE and Applying Evidence to Answer Clinical Questions," *Journal of the American Medical Informatics Association*, 9: 283-293.

Hersh, W., Elliot, D., et al. (1994). "Towards New Measures of Information Retrieval Evaluation," in *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*, Washington, DC. Hanley & Belfus. 895-899.

Hersh, W. and Hickam, D. (1998). "How Well Do Physicians Use Electronic Information Retrieval Systems? A Framework for Investigation and Review of the Literature," *Journal of the American Medical Association*, 280: 1347-1352, http://jama.ama-assn.org/cgi/content/full/280/15/1347.

Hersh, W., Pentecost, J., et al. (1996). "A Task-oriented Approach to Information Retrieval Evaluation," *Journal of the American Society for Information Science*, 47: 50-56.

Hersh, W. and Rindfleisch, T. (2000). "Electronic Publishing of Scholarly Communication in the Biomedical Sciences," *Journal of the American Medical Informatics Association*, 7: 324-325.

Humphreys, B. (2000). "Electronic Health Record Meets Digital Library: A New Environment for Achieving an Old Goal," *Journal of the American Medical Informatics Association*, 7: 444-452.

Lagoze, C. and VandeSompel, H. (2001). "The Open Archives Initiative: Building a Low-barrier Interoperability Framework," in *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*, Roanoke, VA. ACM Press. 54-62.

Lassila, O., Hendler, J., et al. (2001). "The Semantic Web," *Scientific American*, 284(5): 34-43, http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2.

Lawrence, S. (2001). "Online or Invisible?" *Nature*, 411: 521.

Lawrence, S., Giles, C., et al. (1999). "Digital Libraries and Autonomous Citation Indexing," *Computer*, 32: 67-71.

Lesk, M. (1997). *Practical Digital Libraries - Books, Bytes, and Bucks*. San Francisco: Morgan Kaufmann.

McCook, A. (2004). "Open Access to US Govt Work Urged," The Scientist, http://www.biomedcentral.com/news/20040721/01.

McKibbon, K., Haynes, R., et al. (1990). "How Good Are Clinical MEDLINE Searches? A Comparative Study of Clinical End-user and Librarian Searches," *Computers and Biomedical Research*, 23(6): 583-593.

McKiernan, G. (2003. "Open Archives Initiative Service Providers. Part I: Science and Technology," *Library Hi Tech News*, 20(9): 30-38, http://www.public.iastate.edu/~gerrymck/OAI-SP-I.pdf.

McLellan, F. (2003). "US Bill Says Government Funded Work Must Be Open Access," *Lancet*, 362: 52.

Meek, J. (2001). "Science World in Revolt at Power of the Journal Owners," *The Guardian*, http://www.guardian.co.uk/Archive/Article/0,4273,4193292,00.html.

Miller, E. (1998). "An Introduction to the Resource Description Framework," *D-Lib Magazine*, 4, http://www.dlib.org/dlib/may98/miller/05miller.html.

Miller, N., Lacroix, E., et al. (2000). "MEDLINEplus: Building and Maintaining the National Library of Medicine's Consumer Health Web Service," B*ulletin of the Medical Library Association*, 88: 11-17.

Mynatt, B., Leventhal, L., et al. (1992). "Hypertext or Book: Which Is Better for Answering Questions?" in *Proceedings of Computer-Human Interface 92*. 19-25.

Paskin, N. (1999). "DOI: Current Status and Outlook," *D-Lib Magazine*, 5, http://www.dlib.org/dlib/may99/05paskin.html.

Perkel, J. (2003). "Feeding the Info Junkies," *The Scientist*. June 2, 2003. 39, http://www.the-scientist.com/yr2003/jun/feature14_030602.html.

Ponte, J. and Croft, W. (1998). "A Language Modeling Approach to Information Retrieval," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia. ACM Press. 275-281.

Pratt, W., Hearst, M., et al. (1999). "A Knowledge-based Approach to Organizing Retrieved Documents," in *Proceedings of the 16th National Conference on Artificial Intelligence*, Orlando, FL. AAAI. 80-85.

Redman, P., Kelly, J., et al. (1997). "Common Ground: The HealthWeb Project as a Model for Internet Collaboration," B*ulletin of the Medical Library Association*, 85: 325-330.

Robertson, S. and Walker, S. (1994). "Some Simple Effective Aproximations to the 2-Poisson Model for Probabilistic Weighted Retrieval," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland. Springer-Verlag. 232-241.

Salton, G. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Salton, G. (1991). "Developments in Automatic Text Retrieval," *Science*, 253: 974-980.
Salton, G., Fox, E., et al. (1983). "Extended Boolean Information Retrieval," *Communications of the ACM*, 26: 1022-1036.
Salton, G. and Lesk, M. (1965). "The SMART Automatic Document Retrieval System: An Illustration," *Communications of the ACM*, 8: 391-398.
Singhal, A., Buckley, C., et al. (1996). "Pivoted Document Length Normalization," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland. ACM Press. 21-29.
Sollins, K. and Masinter, L. (1994). Functional Requirements for Uniform Resource Names. Internet Engineering Task Force, http://www.w3.org/Addressing/rfc1737.txt.
Spitzer, V., Ackerman, M., et al. (1996). "The Visible Human Male: A Technical Report.," *Journal of the American Medical Informatics Association*, 3: 118-130.
Srinivasan, P. (1996). "Query Expansion and MEDLINE," *Information Processing and Management*, 32: 431-444.
Swanson, D. (1988). "Historical Note: Information Retrieval and the Future of an Illusion," *Journal of the American Society for Information Science*, 39: 92-98.
Tibbo, H. (2001). "Archival Perspectives on the Emerging Digital Library," *Communications of the ACM*, 44(5): 69-70.
vanRijsbergen, C. (1979). *Information Retrieval*. London. Butterworth.
Voorhees, E. (1998). "Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia. ACM Press. 315-323.
Voorhees, E. and Harman, D. (2000). "Overview of the Sixth Text REtrieval Conference (TREC)," *Information Processing and Management*, 36: 3-36.
Voorhees, E. and Harman, D. (2001). "Overview of TREC 2001," in *Proceedings of the Text Retrieval Conference 2001*, Gaithersburg, MD. 1-15.
Weibel, S. (1996). "The Dublin Core: A Simple Content Description Model for Electronic Resources," *ASIS Bulletin*, 24(1): 9-11, http://www.asis.org/Bulletin/Oct-97/weibel.htm.
Wildemuth, B., DeBliek, R., et al. (1995). "Medical Students' Personal Knowledge, Searching Proficiency, and Database Use in Problem Solving," *Journal of the American Society for Information Science*, 46: 590-607.
Xu, J. and Croft, W. (1996). "Query Expansion Using Local and Global Document Analysis," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland. ACM Press. 4-11.

SUGGESTED READINGS

Baeza-Yates, R. and Ribeiro-Neto, B., eds. 1999. *Modern Information Retrieval*. New York. McGraw-Hill. A book surveying most of the automated approaches to information retrieval.

Frakes, W.B., Baeza-Yates, R. *Information Retrieval: Data Structures and Algorithms*, Englewood Cliffs, NJ: Prentice-Hall, 1992. A textbook on implementation of information retrieval systems. Covers all of the major data structures and algorithms, including inverted files, ranking algorithms, stop word lists, and stemming. There are plentiful examples of code in the C programming language.

Hersh, W.R. *Information Retrieval, A Health and Biomedical Perspective* (Second Edition), New York: Springer-Verlag, 2003. A textbook on information retrieval systems in the health and biomedical domain that covers the state of the art as well as research systems.

Humphreys, B., Lindberg, D., et al. 1998. *The Unified Medical Language System: an informatics research collaboration*. Journal of the American Medical Informatics Association, 5: 1-11. A paper describing the motivation and implementation of the National Library of Medicine's Unified Medical Language System.

Miles, W.D. *A History of the National Library of Medicine*, Bethesda, MD: U.S. Dept. of Health and Human Services, 1982. A comprehensive history of the National Library of Medicine and its forerunners, covering the story of Dr. John Shaw Billings and his founding of Index Medicus to the modern implementation of MEDLINE.

Salton, G. Developments in automatic text retrieval, *Science*, 253: 974-980, 1991. The last succinct exposition of word-statistical retrieval systems from the person who originated the approach.

ONLINE RESOURCES

Biomed Central
http://www.biomedcentral.com

Highwire Press
http://www.highwire.org

National Center for Biotechnology Information
http://www.ncbi.nlm.nih.gov

National Library of Medicine
http://www.nlm.nih.gov

ACM Digital Library
http://www.acm.org/dl

CiteSeer
http://citeseer.ist.psu.edu

D-Lib Magazine
http://www.dlib.org

MEDLINEplus consumer health information resource
http://medlineplus.gov

PubMed access to MEDLINE
http://pubmed.gov

TREC
http://trec.nist.gov

QUESTIONS FOR DISCUSSION

1. With the advent of full-text searching, should the National Library of Medicine abandon human indexing of citations in MEDLINE? Why or why not?

2. Explain why open access publishing is or is not a good idea.

3. Devise a curriculum for teaching clinicians, researchers, or patients the most important points about searching for health-related information.

4. What are the limitations of recall and precision as evaluation measures and what alternatives would improve upon them?

5. Describe how one might devise a system that achieved a happy medium between protection of intellectual property and barrier-free access to the archive of science.

6. How might IR systems be developed to lower the effort it takes for clinicians to get to the information they need rapidly in the busy clinical setting?

7. Can standards be developed for digital libraries that facilitate interoperability but maintain ease of use, protection of intellectual property, and long-term preservation of the archive of science?