Information Retrieval and Digital Libraries

21

William R. Hersh

After reading this chapter, you should know the answers to these questions:

- What types of online knowledge-based information are available and useful to clinicians, biomedical researchers, and consumers?
- What are the major components of the information retrieval process?
- What are the major categories of available knowledge-based information?
- How do techniques differ for indexing various types of knowledge-based biomedical information?
- What are the major approaches to retrieval of knowledge-based biomedical information?
- How effectively do searchers use information retrieval systems?
- What are the important research directions in information retrieval?
- What are the major challenges to making digital libraries effective for health and biomedical users?

Information retrieval (**IR**), sometimes called **search**, is the field concerned with the acquisition, organization, and searching of knowledge-based information (Hersh 2009). Although biomedical IR has traditionally concentrated on the retrieval of text from the biomedical literature, the domain over which IR can be effectively applied has broadened considerably with the advent of multimedia publishing and vast storehouses of images, video, chemical structures, gene and protein sequences, and a wide range of other digital media of relevance to biomedical education, research, and patient care. With the proliferation of IR systems and online content, the notion of the library has changed substantially, and new digital libraries have emerged (Lindberg and Humphreys 2005).

IR systems and digital libraries store and disseminate knowledge-based information. What exactly does that mean? Although there are many ways to classify biomedical information and the informatics applications that use them, in this chapter we will broadly divide them into two categories. *Patient-specific information* applies to individual patients. Its purpose is to tell health care providers, administrators, and researchers about the health and disease of a patient. This information comprises the patient's medical record. The second category of biomedical information is *knowledge-based information*. This is information that has been derived and organized from observational or experimental research. In the case of clinical research, this information provides clinicians, administrators, and researchers with knowledge derived from experiments and observations, which can then be applied to

W.R. Hersh, MD, FACMI, FACP
Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, 3181 SW Sam Jackson Park Rd., Mail Code BICC, Portland 97239, OR, USA
e-mail: hersh@ohsu.edu

This chapter is adapted from an earlier version in the third edition authored by William R. Hersh, P. Zoë Stavri, and William M. Detmer.

E.H. Shortliffe, J.J. Cimino (eds.), *Biomedical Informatics*,
DOI 10.1007/978-1-4471-4474-8_21, © Springer-Verlag London 2014

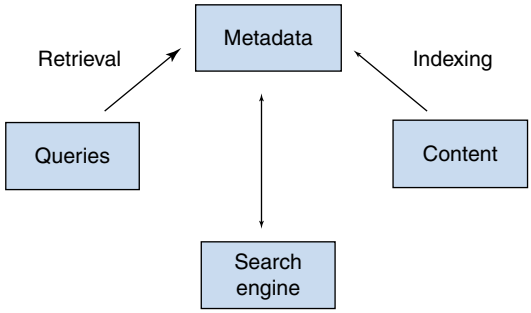

Fig. 21.1 Basic overview of the information retrieval process. Retrieval is made possible via metadata, which is produced via indexing and applied in queries by users. The metadata is used by the search engine, which directs the user to the content (Reproduced with permission of Springer (Hersh 2009))

individual patients. This information is most commonly provided in books and journals but can take a wide variety of other forms, including clinical practice guidelines, consumer health literature, Web sites, and so forth.

A basic overview of the IR process is shown in Fig. 21.1 and forms the basis for most of this chapter. The overall goal of IR or search is to find **content** that meets a person's information needs. This is done by posing a *query* to the IR system. A **search engine** matches the query to content items through **metadata**, which is "data about data" that describes the content items (Foulonneau and Riley 2008). There are two intellectual processes of IR. **Indexing** is the process of assigning metadata to content items, while **retrieval** is the process of the user entering his or her query and retrieving content items.

21.1 Evolution of Biomedical Information Retrieval

As with many chapters in this volume, this area has changed substantially over the four editions of this book. In the first edition, this chapter was titled "Bibliographic-Retrieval Systems," reflecting the type of knowledge that was accessible at the time. The second edition saw the emergence of the **World Wide Web** (**WWW** or **Web**) as a delivery mechanism for knowledge-based information. In the third edition, we added "Digital Libraries" to the chapter name, reflecting that the entire biomedical library and beyond was now part of available online knowledge. This fourth edition reflects the fact that information is ubiquitous on computers, smartphones, tablets, and other devices.

Although this chapter focuses on the use of computers to facilitate IR, methods for finding and retrieving information from medical sources have been in existence for over a century. In 1879 Dr. John Shaw Billings created Index Medicus to help medical professionals find relevant journal articles (DeBakey 1991). Journal article citations were indexed by author name(s) and subject heading(s) and then aggregated in bound volumes. A scientist or practitioner seeking an article on a topic could manually search the index for the single best-matching subject heading and then be directed to citations of published articles.

The printed **Index Medicus** served as the main biomedical IR source but, in 1966, the National Library of Medicine (NLM) unveiled an electronic version, the **Medical Literature Analysis and Retrieval System** (**MEDLARS**) (Miles 1982). Because computing power and disk storage were very limited, MEDLARS and its follow-on **MEDLARS Online** (**MEDLINE**), stored only limited information for each article, such as author name(s), article title, journal source, and publication date. In addition, the NLM assigned to each article a number of terms from its **Medical Subject Headings** (**MeSH**) vocabulary (see Chap. 7). Searching was done by users having to mail a paper search form to the NLM and receiving results back a few weeks later. Only librarians who had completed a specialized course were allowed to submit searches.

As computing power grew and disk storage became more plentiful in the 1980s, full-text databases began to emerge. These new databases allowed searching of the entire text of medical documents. Although lacking graphics, images, and tables from the original source, these databases made it possible to retrieve the full text of important documents quickly from remote locations. Likewise, with the growth of **time-sharing networks**, end users were now allowed to search the databases directly, though at a substantial cost.

In the early 1990s, the pace of change in the IR field quickened. The advent of the Web and the exponentially increasing power of computers and networks brought a world where vast quantities of medical information from multiple sources with various media extensions were now available over the global Internet (Berners-Lee et al. 1994). In the late 1990s, the NLM made all of its databases available to the entire world for free. Also during this time, the notion of digital libraries developed, with the recognition that the entire array of knowledge-based information could be accessed using this technology (Borgman 1999).

Now in the twenty-first century, use of IR systems and digital libraries has become ubiquitous. Estimates vary, but among individuals who use the Internet in the United States, over 80 % have used it to search for information relevant to their own health or that of an acquaintance (Fox 2011; Taylor 2010). Virtually all physicians use the Internet (Davies 2010). Furthermore, access to systems has gone beyond the traditional personal computer and extended to new devices, such as smartphones and tablet devices.

21.2 Knowledge-Based Information in Health and Biomedicine

Knowledge-based information can be subdivided into two categories. **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 reanalysis of data (e.g., systematic reviews and meta-analyses). **Secondary knowledge–based information** consists of the writing that reviews, condenses, and/or synthesizes the primary literature. The most common examples of this type of literature are books, monographs, and review articles. Secondary literature includes the growing quality of patient/consumer-oriented health information that is increasingly available via the Web. It also encompasses opinion-based writing (such as editorials and position or policy papers), clinical practice guidelines, narrative reviews, and health information on Web pages. In addition, it includes the plethora of pocket-sized manuals that were formerly a staple for practitioners in many professional fields. As will be seen later, secondary literature is the most common type of literature used by physicians.

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 get assistance with seeking information, including information on computers
- Providing work or study space (particularly in universities)

Digital libraries provide some of the same services, but their focus tends to be on the digital aspects of content.

21.2.1 Information Needs and Information Seeking

Different users of knowledge-based information have differing needs based on the nature of their information need and available resources. The information needs and information seeking of physicians have been most extensively studied. Gorman and Helfand (1995) has defined four states of **information need** in the clinical context:

- Unrecognized need—clinician unaware of information need or knowledge deficit
- Recognized need—clinician aware of need but may or may not pursue it
- Pursued need—information seeking occurs but may or may not be successful
- Satisfied need—information seeking successful

There is a great deal of evidence that the majority of information needs are not being satisfied and that IR applications may help. Among the reasons that physicians do not adhere to the

most up-to-date clinical practices is that they often do not recognize that their knowledge is incomplete. While this is not the only reason for such practices, the evidence is compelling. For example, physicians do not provide patients with most up-to-date care (McGlynn et al. 2003), do not adhere to established guidelines (Diamond and Kaul 2008), and vary widely in how they provide care (Wennberg 2010).

When physicians recognize an information need, they are likely to pursue only a minority of unanswered questions. A variety of studies over several decades have demonstrated that physicians in practice have unmet information on the order of two questions for every three patients seen and only pursue answers for about 30 % of these questions (Covell et al. 1985; Ely et al. 1999; Gorman and Helfand 1995). When answers to questions are actually pursued, these studies showed that the most frequent source for answers to questions was colleagues, followed by paper-based textbooks. Therefore, it is not surprising that barriers to satisfying information needs remain (Ely et al. 2002). Physicians use electronic sources more now than were measured in these earlier studies, with the widespread use of the **electronic health record** (**EHR**) as well as ubiquity of portable smartphones and tablets. One possible approach to lowering the barrier to knowledge-based information is to link it more directly with the context of the patient in the EHR (Cimino and delFiol 2007).

The information needs of other users are less well-studied. For consumers, ongoing surveys find about 80 % of all Internet users have searched for personal health information (Fox 2011; Taylor 2010). About 4.5 % of all queries to Web search engines are health-related (Eysenbach and Kohler 2004). The most common type of search focuses on a specific disease or medical problem (66 % of all who have searched), followed by a specific medical treatment or procedure (56 %). Consumers also use the Web to search for physicians, health care institutions, and health insurance. Less studied are the information needs of researchers, but one recurrent finding is the idiosyncratic nature of their use of IR and other systems (Bartlett and Toms 2005).

21.2.2 Changes in Publishing

The Internet and the Web have had a profound impact on the publishing of knowledge-based information. The technical impediments to electronic publishing of journals have been overcome, such that virtually all scientific journals are published electronically now. 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,[1] a spin-off from Stanford University, which has an infrastructure that supports the tasks of 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. When available in electronic form, journal content is easier and more convenient to access. Furthermore, since most scientists have the desire for widespread dissemination of their work, they have incentive for their papers to be available electronically. Not only is there the increased convenience of redistributing reprints, but research has found that freely available on the Web have a higher likelihood of being cited by other papers than those that are not (Eysenbach 2006; Lawrence 2001; Moed 2007). 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; Sox 2009). Printing and mailing, tasks no longer needed in electronic publishing, comprised a significant part of the "added value" from publishers of journals. There is still however 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 currently exist today, were to vanish, there would still be some cost to the production of journals. Thus, while the cost of producing

[1] www.highwire.org

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 (Sox 2009). This introduces some political issues as well. One of them centers around 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 other 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. This problem is exacerbated by the general decline in funding for libraries.

Some proposed models of "open access" scholarly publishing keep the archive of science freely available (Albert 2006; Björk et al. 2010). The basic principle of open access publishing is that authors and/or their institutions pay the cost of production of manuscripts up front after they are accepted through a peer review process. After the paper is published, it becomes freely available on the Web. Since most research is usually funded by grants, the cost of open access publishing should be included in grant budgets. The uptake of publishers adhering to the open access model has been modest, with the most prominent being **Biomed Central** (**BMC**)[2] and the **Public Library of Science** (**PLoS**)[3].

Another model that has emerged is **PubMed Central** (**PMC**, pubmedcentral.gov). PMC is a repository of life science research that provides free access while allowing publishers to maintain copyright and even optionally keep the papers housed 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 National Institutes of Health (NIH)[4] now requires all research funded by its grants to be submitted to PMC, either in the form published by publishers or as a PDF of the last manuscript prior to journal acceptance.[5] Publishers have expressed concern that copyrights give journals more control over the integrity of the papers they publish (Drazen and Curfman 2004). An alternative approach, advocated by non-commercial (professional society) publishers is the **Washington DC Principles for Free Access to Science**,[6] which advocates:

- Reinvestment of revenues in support of science.
- Use of open archives such as PMC as allowed by business constraints.
- Commitment to some free publication, access by low-income countries, and no charges to publish.

21.2.3 Quality of Information

The growth of the Internet and the Web has led to another concern, which is the quality of information available. A large fraction of Web-based health information is aimed at nonprofessional audiences. Many laud this development as empowering those most directly affected by health care—those who consume it (Eysenbach et al. 1999). Others express concern about patients misunderstanding or being purposely misled by incorrect or inappropriately interpreted information (Jadad 1999). Some clinicians also lament the time required to go through stacks of printouts downloaded by patients and brought to the office. The Web is inherently democratic, allowing anyone to post information. This is an asset in a democratic society like that of the United States. However, it is potentially at odds with the operation of a professional field, particularly one like health care, where practitioners are ethically bound and legally required to adhere to the highest standard of

[2] www.biomedcentral.com

[3] www.plos.org

[4] www.nih.gov

[5] www.publicaccess.nih.gov

[6] www.dcprinciples.org

care. Thus, a major concern with health information on the Web is the presence of inaccurate or out-of-date information. Although research in this area has declined, one **systematic review** of studies assessing the quality of health information found that 55 of 79 studies came to the conclusion that quality of information was a problem (Eysenbach et al. 2002).

The impact of poor-quality information is unclear. People have been harmed by incorrect and misleading health information since time immemorial. One well-known self-help expert has argued that patients and consumers actually are savvy enough to understand the limits of quality of information on the Web. This view holds that patients and consumers should be trusted to discern quality using their own abilities to consult different sources of information and to communicate with health care practitioners and with others who share their condition(s) (Ferguson 2002). Indeed, the ideal situation may be a partnership among patients and their health care practitioners, as it has been shown that patients desire that their practitioners be the primary source of recommendations for online information (Tang et al. 1997).

This concern about quality of information has led a number of individuals and organizations to develop guidelines for assessing the quality of health information. These guidelines usually have explicit criteria for a Web page that a reader can apply to determine whether a potential source of information has attributes consistent with high quality. One of the earliest and most widely quoted set of criteria was published in JAMA (Silberg et al. 1997). These criteria stated that Web pages should contain the name, affiliation, and credentials of the author; references to the claims made; explicit listing of any perceived or real conflict of interest; and date of most recent update. Another early set of criteria was the **Health on the Net** (HON) codes,[7] a set of voluntary codes of conduct for health-related Web sites. Sites that adhere to the HON codes can display the HON logo. Another approach to insuring Web site quality is accreditation by a third party. **URAC** (formerly, the Utilization Review Accreditation Commission) has a process for such accreditation.[8] The URAC standards cover six general issues: health content editorial process, disclosure of financial relationships, linking to other Web sites, privacy and security, consumer complaint mechanisms, and internal processes required to maintain quality over time.

21.2.4 Evidence-Based Medicine

The growing quantity of clinical information available in IR systems and digital libraries requires new approaches to select that which is best to use for clinical decisions. The philosophy guiding this approach is **evidence-based medicine** (EBM), which can be viewed a set of tools to inform clinical decision making. It allows clinical experience ("art") to be integrated with best clinical science (Haynes et al. 2002). Also, EBM makes the medical literature more clinically applicable and relevant. In addition, it requires the user to be facile with computers and IR systems. There are many well-known books and Web resources for EBM, with the original textbook now in a fourth edition (Straus et al. 2005). The process of EBM involves three general steps:

- Phrasing a clinical question that is pertinent and answerable.
- Identifying evidence (studies in articles) that address the question.
- Critically appraising the evidence to determine whether it applies to the patient.

The phrasing of the clinical question is an often-overlooked portion of the EBM process. There are two general types of clinical question: background questions and foreground questions (Straus et al. 2005). **Background questions** ask for general knowledge about a disorder, whereas **foreground questions** ask for knowledge about managing patients with a disorder. Background questions are generally best answered with textbooks and classical review articles, whereas foreground questions are answered using EBM

[7] www.hon.ch

[8] http://www.urac.org/consumers/overview.aspx

techniques. There are four major foreground question categories:

- Therapy (or intervention)—benefit of treatment or prevention.
- Diagnosis—test diagnosing disease.
- Harm—detrimental health effects of a disease, environmental exposure (natural or man-made), or medical intervention.
- Prognosis—outcome of disease course.

Identifying evidence involves selecting the best evidence for a given type of question. EBM proponents advocate, for example, that randomized controlled trials or a meta-analysis that combines multiple trials provide the best evidence for or against particular health care interventions. Likewise, diagnostic test accuracy is best assessed with comparison to a known gold standard in an appropriate spectrum of patients to whom the test will be applied (see Chap. 3). Questions of harm can be answered by randomized controlled trials when it is ethical to do so; otherwise they are best answered with observational case control or cohort studies. There are checklists of attributes for these different types of studies that allow their critical appraisal and applicability to a given patient in the EBM resources described above.

The original approach to EBM has evolved over time, with less emphasis on critically appraising original evidence and more on synthesized evidence being made readily available to clinicians, usually through electronic sources, including clinical decision support systems (see Chap. 22, DiCenso et al. 2009; Hersh 1999). There have also been a number of criticisms of EBM, as summarized by Cohen et al. (2004).

21.3 Content of Knowledge-Based Information Resources

The previous sections of this chapter have described some of the issues and concerns surrounding the production and use of knowledge-based information in biomedicine. It is useful to classify the information to gain a better understanding of its structure and function. In this section, we classify content into bibliographic, full-text, annotated, and aggregated categories, although some content does not neatly fit within them.

21.3.1 Bibliographic Content

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 contains bibliographic references to all of the biomedical articles, editorials, and letters to the editors in approximately 5,000 scientific journals. The journals are chosen for inclusion by an advisory committee of subject experts convened by NIH. At present, about 700,000 references are added to MEDLINE yearly. It contained over 22 million references by the end of 2012.

The current MEDLINE record contains up to 49 fields. A clinician may be interested in just a handful of these fields, such as the title, abstract, and indexing terms. But other fields contain specific information that may be of great importance to other audiences. For example, a genome researcher might be highly interested in the Supplementary Information (SI) field to link to genomic databases. Even the clinician may, however, derive benefit from some of the other fields. For example, the Publication Type (PT) field can help in the application of EBM, such as when one is searching for a practice guideline or a randomized controlled trial. MEDLINE is accessible by many means and available without charge via the **PubMed** system (pubmed.gov), produced by the **National Center for Biotechnology Information** (**NCBI**,)[9] of the NLM. A number of other information vendors, such as Ovid Technologies[10] and Aries Systems,[11] license the content of MEDLINE and other databases and

[9] www.ncbi.nlm.nih.gov

[10] www.ovid.com

[11] www.ariessys.com

provide value-added services that can be accessed for a fee by individuals and institutions.

MEDLINE is only one of many databases produced by the NLM. Other more specialized databases are also available, covering topics from AIDS to space medicine and toxicology. There are several 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 the **Cumulative Index to Nursing and Allied Health Literature** (CINAHL, CINAHL Information Systems),[12] which covers nursing and allied health literature, including physical therapy, occupational therapy, laboratory technology, health education, physician assistants, and medical records.

Another well-known bibliographic databases is **EMBASE**,[13] which is sometimes referred to as the "European MEDLINE." It contains over 24 million records and covers many of the same medical journals as MEDLINE but with a more international focus, including more non-English-language journals. These journals are often important for those carrying out meta-analyses and systematic reviews, which need access to all the studies done across the world.

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; see Sects. 21.3 and 21.4, below). 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:

- 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.
- HON Select[14] —a European catalog of quality-filtered, clinician-oriented Web content from the HON foundation.
- Translating Research into Practice (TRIP)[15]—a database of content deemed to meet high standards of EBM.
- Open Directory[16] —a general Web catalog that has significant health content.

Another more modern bibliographic resource is the **National Guidelines Clearinghouse** (**NGC**)[17]. Produced by the Agency for Health care Research and Quality (AHRQ), it contains exhaustive information about clinical practice guidelines. Some of the guidelines produced are freely available, published electronically and/or on paper. Others are proprietary, in which case a link is provided to a location at which the guideline can be ordered or purchased. The overall goal of the NGC is to make evidence-based clinical practice guidelines and related abstract, summary, and comparison materials widely available to health care and other professionals.

A final kind of bibliographic-like content consists of **RSS feeds** (originally RDF Site Summary, often dubbed "Really Simple Syndication"), which are short summaries of Web content, typically news, journal articles, blog postings, and other content. Users set up an RSS aggregator, which can be though a Web browser, email client, or standalone software, configured for the RSS feed desired, with an option to add a filter for specific content. There are two versions of RSS (1.0 and 2.0) but both provide:

- Title—name of item
- Link—URL to content
- Description—a brief description of the content

21.3.2 Full-text Content

The second type of content is **full-text content**. A large component of this content consists of the online versions of books and periodicals. As already noted, most traditionally paper-based medical literature, from textbooks to journals, is now available electronically. The electronic versions

[12] http://www.ebscohost.com/cinahl/
[13] www.embase.com
[14] www.hon.ch/HONselect
[15] www.tripdatabase.com
[16] www.dmoz.org
[17] www.guideline.gov

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 the vast number of static and dynamic Web pages at a discrete Web location.

Electronic publication of journals allows additional features not possible in the print world. Journal editors often clash with authors over the length of published papers (editors want them short for readability whereas authors want them long to be able to present all ideas and results). To address this situation, the British Medical Journal (BMJ) initiated an **electronic-long**, **paper-short** (**ELPS**) system that provides on the Web site supplemental material that did not appear in the print version of the journal. Journal Web sites can provide supplementary data of results, images, and even raw data. A journal Web site also allows more dialog 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. PubMed maintains a field for the Web address of the full-text paper. This linkage is active when the PubMed record is displayed, but users may be met by a password screen if the article is not available for free. Many sites allow both access to subscribers or a pay-per-view facility. Many academic organizations now maintain large numbers of subscriptions to journals available to faculty, staff, and students. Other publishers, such as Ovid and MD Consult,[18] provide access within their own password-protected interfaces to articles from journals that they have licensed for use in their systems.

The most common secondary literature source is traditional textbooks, an increasing number of which are available in computer form. A common approach with textbooks is bundling them, sometimes with linkages across the bundled texts. An early bundler of textbooks was Stat!-Ref (Teton Data Systems)[19] that, like many, began as a CD-ROM product and then moved to the Web. Stat!-Ref offers over 30 textbooks. Most other publishers have similar aggregated their libraries of textbooks and other content. Another collection of textbooks is the NCBI Bookshelf, which contains many volumes on biomedical research topics.[20] A separate book on the NCBI Web site is Online Mendelian Inheritance in Man (OMIM)[21], which is continually updated with new information about the genomic causes of human disease.

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. Finally, 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 2–5 years.

As noted above, Web sites are another form of full-text information. Probably the most effective provider of Web-based health information is the U.S. government. Not only do they produce bibliographic databases, but the NLM, AHRQ, the National Cancer Institute (NCI), Centers for Disease Control (CDC), and others have also been innovative in providing comprehensive full-text information for health care providers and consumers. One example is the popular CDC Travel site.[22] Some of these will be described later as aggregations, since they provide many different types of resources.

[18] www.mdconsult.com

[19] www.statref.com

[20] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books

[21] http://www.ncbi.nlm.nih.gov/omim

[22] http://www.cdc.gov/travel/

A large number of commercial biomedical and health Web sites have emerged in recent years. On the consumer side, they include more than just collections of text; they also include interaction with experts, online stores, and catalogs of links to other sites. Among the best known of these are Intelihealth[23] and NetWellness.[24] There are also Web sites, either from medical societies or companies, that provide information geared toward health care providers, typically overviews of diseases, their diagnosis, and treatment; medical news and other resources for providers are often offered as well.

Other sources of on-line health-related content include encyclopedias, the so-called "**body of knowledge**" (BOK; the complete set of concepts, terms and activities that make up a professional domain), and **Weblogs** or **blogs**. A well-known online encyclopedia with a great deal of health-related information is Wikipedia,[25] which features a distributed authorship process whose content has been found to reliable (Giles 2005; Nicholson 2006) and frequently shows up near the top in health-related Web searches (Laurent and Vickers 2009). A growing number of organizations have a body of knowledge, such as the **American Health Information Management Association** (**AHIMA**).[26] Blogs tend to carry a stream of consciousness but often high-quality information is posted within them.

21.3.3 Annotated Content

The third category consists of **annotated 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 genomic research
- **Citation databases**—bibliographic linkages of scientific literature
- **EBM databases**—highly structured collections of clinical evidence
- Other databases—miscellaneous other collections

A great number of biomedical image databases are available on the Web. These include:

- Visible Human[27]
- BrighamRad[28]
- WebPath[29]
- Pathology Education Instructional Resource (PEIR)[30]
- DermIS[31]
- VisualDX[32]

Many genomics databases are available on the Web. The first issue each year of the journal *Nucleic Acids Research* (NAR) catalogs and describes these databases, and is now available by open access means (Galperin and Cochrane 2011). NAR also maintains an ongoing database of such databases, the Molecular Biology Database Collection.[33] Among the most important of these databases are those available from NCBI (Sayers et al. 2011). All their databases are linked among themselves, along with PubMed and OMIM, and are searchable via the **Entrez** system.[34] More details on the specific content of genomics databases is provided in Chap. 24.

Citation databases provide linkages to articles that cite others across the scientific literature. The earliest citation databases were the Science Citation Index (SCI, Thomspon-Reuters) and Social Science Citation Index (SSCI, Thomspon-Reuters), which are now part of the larger Web of

[23] www.intelihealth.com
[24] www.netwellness.com
[25] www.wikipedia.org
[26] http://library.ahima.org/bok
[27] http://www.nlm.nih.gov/research/visible/visible_human.html
[28] http://brighamrad.harvard.edu/
[29] http://library.med.utah.edu/WebPath/webpath.html
[30] www.peir.net
[31] www.dermis.net
[32] www.visualdx.com (requires a subscription fee)
[33] http://www.oxfordjournals.org/nar/database/a/
[34] www.ncbi.nlm.nih.gov/Entrez

Science. Two well-known bibliographic databases for biomedical and health topics that also have citation links include SCOPUS[35] and Google Scholar.[36] These three were recently compared for their features and coverage (Kulkarni et al. 2009). A final citation database of note is CiteSeer,[37] which focuses on computer and information science, including biomedical informatics.

EBM databases are devoted to providing annotated evidence-based information. Some examples (all available with through subscription fees) include:

- The Cochrane Database of Systematic Reviews—one of the original collections of systematic reviews[38]
- Clinical Evidence—an "evidence formulary"[39]
- Up-to-Date—content centered around clinical questions[40]
- InfoPOEMS—"patient-oriented evidence that matters"[41]
- Physicians' Information and Education Resource (PIER)—"practice guidance statements" for which every test and treatment has associated ratings of the evidence to support them[42]

There is a growing market for a related type of evidence-based content in the form of clinical decision support order sets, rules, and health/disease management templates. Publishers include EHR vendors whose systems employ this content as well as other vendors such as Zynx[43] and Thomson-Reuters.[44]

There are a variety of other annotated content. The ClinicalTrials.gov database began as a database of clinical trials sponsored by NIH. In recent years it has expanded its scope to be a registry of all clinical trials (DeAngelis et al. 2005; Laine et al. 2007) and to containing actual results of trials (Zarin et al. 2011). Another important database for researchers is NIH RePORTER,[45] which is a database of all research funded by NIH.

21.3.4 Aggregated Content

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 the diverse needs of 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**[46] from the NLM. MedlinePlus includes all of the types of content previously described, aggregated for easy access to a given topic. MedlinePlus contains health topics, drug information, medical dictionaries, directories, and other resources. 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. Most of the major publishers now aggregate all of their content in packages for clinicians. Another aggregated resource for clinicians is **Merck Medicus**,[47] developed by the well-known publisher and pharmaceutical house, is available for free to all licensed U.S. physicians, and includes such well-known resources as Harrison's Online, MDConsult, and DXplain.

[35] www.scopus.com
[36] scholar.google.com
[37] http://citeseerx.ist.psu.edu/
[38] www.cochrane.org
[39] www.clinicalevidence.com
[40] www.uptodate.com
[41] www.infopoems.com
[42] www.pier.acponline.org
[43] www.zynxhealth.com
[44] www.thomsonreuters.com
[45] http://projectreporter.nih.gov/reporter.cfm
[46] www.medlineplus.gov
[47] www.merckmedicus.com

Another well-known group of aggregations of content for genomics researchers is the **model organism databases**. These databases bring together bibliographic databases, full text, and databases of sequences, structure, and function for organisms whose genomic data have been highly characterized. One of the oldest and most developed model organism databases is the Mouse Genome Informatics resource.[48] More details are provided in Chap. 22.

21.4 Indexing

As noted at the beginning of the chapter, indexing is the process of assigning metadata to content to facilitate its retrieval. Most modern commercial content is indexed in two ways:

1. **Manual indexing**—where human indexers, usually using a controlled terminology, assign indexing terms and attributes to documents, often following a specific protocol.
2. **Automated indexing**—where 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 done most commonly with bibliographic databases and annotated content. 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. Before covering these types of indexing in detail, let us first discuss controlled terminologies.

21.4.1 Controlled Terminologies

A **controlled terminology** contains a set of terms that can be applied to a task, such as indexing. When the terminology defines the terms, it is usually called a **vocabulary**. When it contains variants or synonyms of terms, it is also called a **thesaurus**. Before discussing actual terminologies, it is useful to define some terms. A **concept** is an idea or object that exists 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 terminology usually contains a list of terms that are the canonical representations of the concepts. If it is athesaurus, it contains 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 (e.g., MeSH tree explosions that add terms from an entire portion of the hierarchy to augment a search).
- **Synonym**—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 MeSH terminology is used to manually index most of the databases produced by the NLM (Coletti and Bleich 2001). The latest version contains over 26,000 subject headings (the word MeSH uses for the canonical representation of its concepts). It also contains over 170,000 synonyms to those terms, which in MeSH jargon are called **entry terms**. In addition, MeSH contains the three types of relationships described in the previous paragraph:

- Hierarchical—MeSH is organized hierarchically into 16 trees, such as Diseases, Organisms, and Chemicals and Drugs
- Synonym—MeSH contains a vast number of entry terms, which are synonyms of the headings

[48] www.informatics.jax.org

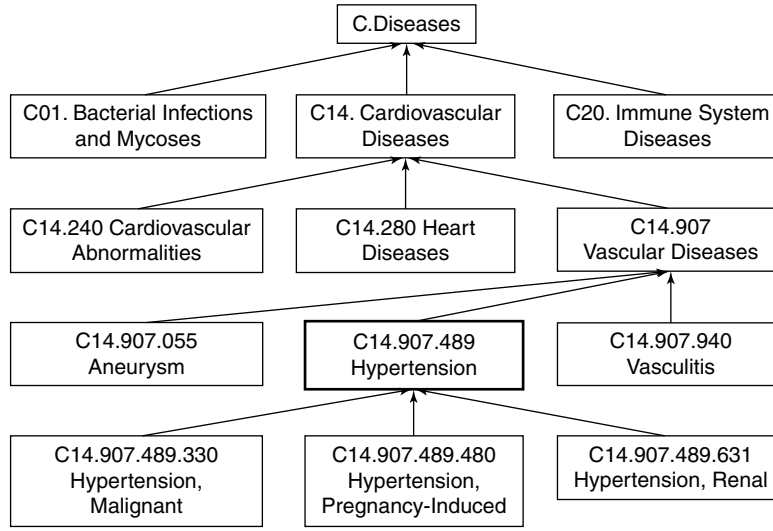

Fig. 21.2 A slice through the Medical Subject Headings (MeSH) hierarchy for "Hypertension" and related terms, showing the location of the term in the C. Diseases. The arrows show links to broader terms in the hierarchy, while the codes give the tree address used internally by the MeSH system (Reproduced with permission of Springer (Hersh 2009))

- Related—terms that may be useful for searchers to add to their searches when appropriate are suggested for many headings

The MeSH terminology files, their associated data, and their supporting documentation are available on the NLM's MeSH Web site.[49] There is also a browser that facilitates exploration of the terminology.[50] Figure 21.2 shows a slice through the MeSH hierarchy for "Hypertension" and related cardiovascular diseases in the C. Diseases tree.

There are features of MeSH designed to assist indexers in making documents more retrievable. One of these is **subheadings**, which are qualifiers of subject headings that narrow the focus of a term. In 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 one particular part of the MeSH hierarchy (called the "Z tree", because their term codes start with "Z"). 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. EMBASE has a terminology called **EMTREE**, which has many features similar to those of MeSH.[51]

One problem with controlled terminologies, not limited to IR systems, is their proliferation. As already described in Chap. 7, there is great need for linkage across these different terminologies. This was the primary motivation for the **Unified Medical Language System (UMLS) Project**, which was undertaken in the 1980s to address this problem (Humphreys et al. 1998).

[49] http://www.nlm.nih.gov/mesh/

[50] http://www.nlm.nih.gov/mesh/MBrowser.html

[51] http://www.embase.com/info/helpfiles/emtree-tool/emtree-thesaurus

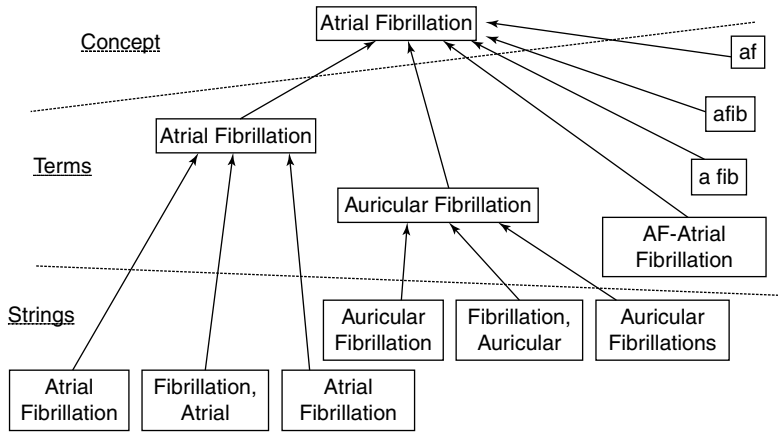

Fig. 21.3 Concepts, terms, and strings for the Metathesaurus concept atrial fibrillation. Each string may occur in more than one vocabulary, in which case each would be an atom (Reproduced with permission of Springer (Hersh 2009))

There are three components of the **UMLS Knowledge Sources**: the **Metathesaurus**, the **UMLS Semantic Network**, and the **Specialist Lexicon**. The Metathesaurus component of the UMLS links parts or all of over 100 terminologies (Bodenreider 2004).

In the Metathesaurus, all terms that are conceptually the same are linked together as a concept. Each concept may have one or more terms, each of which represents an expression of the concept from a source terminology that is not just a simple lexical variant (i.e., differs only in word ending or order). Each term may consist of one or more strings that represent all the lexical variants that are represented for that term in the source terminologies. One of each term's strings is designated as the preferred form, and the preferred string of the preferred term is known as the canonical form of the concept. There are rules of precedence for determining the canonical form, the main one being that the MeSH heading is used if one of the source terminologies for the concept is MeSH.

Each Metathesaurus concept has a single concept unique identifier (CUI). Each term has one term unique identifier (LUI), all of which are linked to the one (or more) CUIs with which they are associated. Likewise, each string has one string unique identifier (SUI), which is likewise linked to the LUIs in which they occur. In addition, each string has an atomic unique identifier (AUI) that represents information from each instance of the string in each vocabulary. Figure 21.3 depicts the English-language concepts, terms, and strings for the Metathesaurus concept atrial fibrillation. (Each string may occur in more than one vocabulary, in which case each would be an atom.) The canonical form of the concept and one of its terms is atrial fibrillation. Within both terms are several strings that vary in word order and case.

The Metathesaurus contains a wealth of additional information. In addition to the synonym relationships between concepts, terms, and strings described earlier, there are also non-synonym relationships between concepts. There are a great many attributes for the concepts, terms, strings, and atoms, such as definitions, lexical types, and occurrence in various data sources. Also provided with the Metathesaurus is a word index that connects each word to all the strings it occurs in, along with its concept, term, string, and atomic identifiers.

21.4.2 Manual Indexing

Manual indexing is most commonly done for bibliographic and annotated content, although it is sometimes for other types of content as well. Manual indexing is usually done by means of a controlled terminology 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, 1983). Subsequent modifications have occurred with changes to MEDLINE, other databases, and MeSH over the years. The major concepts of the article, usually from two to five headings, are designed as main 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. Although MEDLINE indexing is still manual, indexers are aided by a variety of electronic tools for selecting and assigning MeSH terms.

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 Sect. 21.4.3, below). One exception to this is MDConsult,[52] which uses back-of-book indexes to point to specific sections in its online books.

Manual indexing of Web content is challenging. With billions of 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 earlier. These catalogs contain 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 that are often complementary. The first approach, that of applying metadata to Web pages and sites, is exemplified by the **Dublin Core Metadata Initiative** (**DCMI**,)[53] (Weibel and Koch 2000). The second approach, to build directories of content, was popularized initially by the Yahoo search engine.[54] A more open approach to building directories was taken up by the Open Directory Project,[55] which carries on the structuring of the directory and entry of content by volunteers across the world.

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. The specification has defined 15 elements, as shown in Table 21.1. The DCMI was recently approved as a standard by the **National Information Standards Organization** (**NISO**) with the designation Z39.85. It is also a standard with the **International Organization for Standards** (**ISO**), ISO Standard 15836:2009.

There have been some medical adaptations of the DCMI. The most developed of these is the Catalogue et Index des Sites Médicaux Francophones (CISMeF).[56] (Darmoni et al. 2000). A catalog of French-language health resources on the Web, CISMeF has used DCMI to catalog over 40,000 Web pages, including information resources (e.g., practice guidelines, consensus development conferences), organizations (e.g., hospitals, medical schools, pharmaceutical companies), and databases. The Subject field uses the French translation of MeSH but also includes the English translation. For Type, a list of common Web resources has been enumerated.

While Dublin Core Metadata was originally envisioned to be included in **Hypertext Markup Language** (**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**) (Akerkar 2009). A framework for describing and

[52] www.mdconsult.com

[53] www.dublincore.org

[54] www.yahoo.com

[55] www.dmoz.org

[56] www.cismef.org

Table 21.1 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
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

interchanging metadata, RDF is usually expressed in **Extensible Markup Language** (**XML**), a standard for data interchange on the Web. RDF also forms the basis of what some call the future of the Web as a repository not only of content but also of knowledge, which is also referred to as the **Semantic Web** (Akerkar 2009). Dublin Core Metadata (or any type of metadata) can be represented in RDF.

Another approach to manually indexing content on the Web has been to create directories of content. The first major effort to create these was for use in the Yahoo! search engine, which created a subject hierarchy and assigned Web sites to elements within it. When concern began to emerge that the Yahoo directory was proprietary and not necessarily representative of the Web community at large, 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 61–75 %. The least consistent indexing occurred with subheadings, especially those assigned to non-central-concept headings, which had a consistency of less than 35 %. A repeat of this study in more recent times found comparable results (Marcetich et al. 2004). 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 semiautomated means of doing so (Aronson et al. 2004).

21.4.3 Automated Indexing

In automated indexing, the indexing is done by a computer. Although the mechanical running of the automated indexing process lacks cognitive input, considerable intellectual effort may have

gone into development of the system for doing it, so this form of indexing still qualifies as an intellectual process. 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 might 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 term or individual words within it. As will be seen in the next chapter, 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 allowed only word indexing began to emerge. This trend increased with the advent of the Web.

Word indexing is typically done by defining all consecutive **alphanumeric** sequences between white space (which consists of spaces, punctuation, carriage returns, and other non-alphanumeric characters) as words. Systems must take particular care to apply the same process to documents and the user's query, especially with characters such as hyphens and apostrophes. Many 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 word 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 list of 250–500 words more typically used. Examples of the latter are the 250-word list of van Rijsbergen (1979), the 471-word list of Fox (1992), and the PubMed stop list (Anonymous 2007). 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. Both stop word remove and stemming 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:

$$IDF\left(term\right) = \log \frac{\text{number of documents in database}}{\text{number of documents with term}} + 1 \tag{21.1}$$

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:

$$TF\left(term, document\right) = \text{frequency of term in document} \tag{21.2}$$

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

$$WEIGHT\left(term, document\right) = TF\left(term, document\right) * IDF\left(term\right) \tag{21.3}$$

Another automated indexing approach generating increased interest is the use of **link-based** methods, fueled by the success of the **Google** search engine.[57] This approach gives weight to pages based on how often they are cited by other pages. The **PageRank** (**PR**) **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 PR, whereas a more obscure page will have a lower PR. Google has also had to develop new computer architectures and algorithms to maintain pace with indexing the Web, leading to a new paradigm for such large-scale processing called MapRecuce (Dean and Ghemawat 2008; Lin and Dyer 2010).

General-purpose search engines such as Google and Microsoft Bing use word-based approaches and variants of the PageRank algorithm for indexing. They amass the content in their search systems by "crawling" the Web, collecting and indexing every object they find on the Web. This includes not only HTML pages, but other files as well, including Microsoft Word, Portable Document Format (PDF), and images.

Word 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 (CHF) 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.

Chapter 8 on Natural Language Processing (NLP) describes automated methods for addressing these limitations.

21.5 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 to 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.

21.5.1 Exact-Match Retrieval

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

[57] www.google.com

though Salton was developing the partial-match approach in research systems during that time (Salton and McGill 1983). Currently, exact-match searching tends to be associated with retrieval from bibliographic and annotated databases, while the partial-match approach tends to be used with full-text searching.

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 with 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 is applied to a pair of sets, with the result being the documents found in the first set but not in the second 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.

21.5.2 Partial-Match Retrieval

Although **partial-match searching** was conceptualized very early, 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. 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 and McGill (1983), who pioneered the approach in the 1960s. 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., (Salton 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:

$$Document\,weight = \sum_{all\ queryterms} Weight\,of\,term\ in\ query * Weight\,of\,term\ in\,document \quad (21.4)$$

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.)

21.5.3 Retrieval Systems

This section describes searching systems used to retrieve content from the four categories previously described in Sect. 21.3.

As noted above, PubMed is the system at NLM that searches MEDLINE and other bibliographic databases. 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 (described in the on-line help system of PubMed). In this automatic term mapping, the system attempts to map user input, in succession, to MeSH terms, journals 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). A results screen on a search combining the angiotensin-converting (ACE) inhibitor class of drugs and the disease congestive heart failure (CHF) is shown in Fig. 21.4.

PubMed allows the use of wild-card characters. It also allows phrase searching whereby 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 "Limits." These include publication

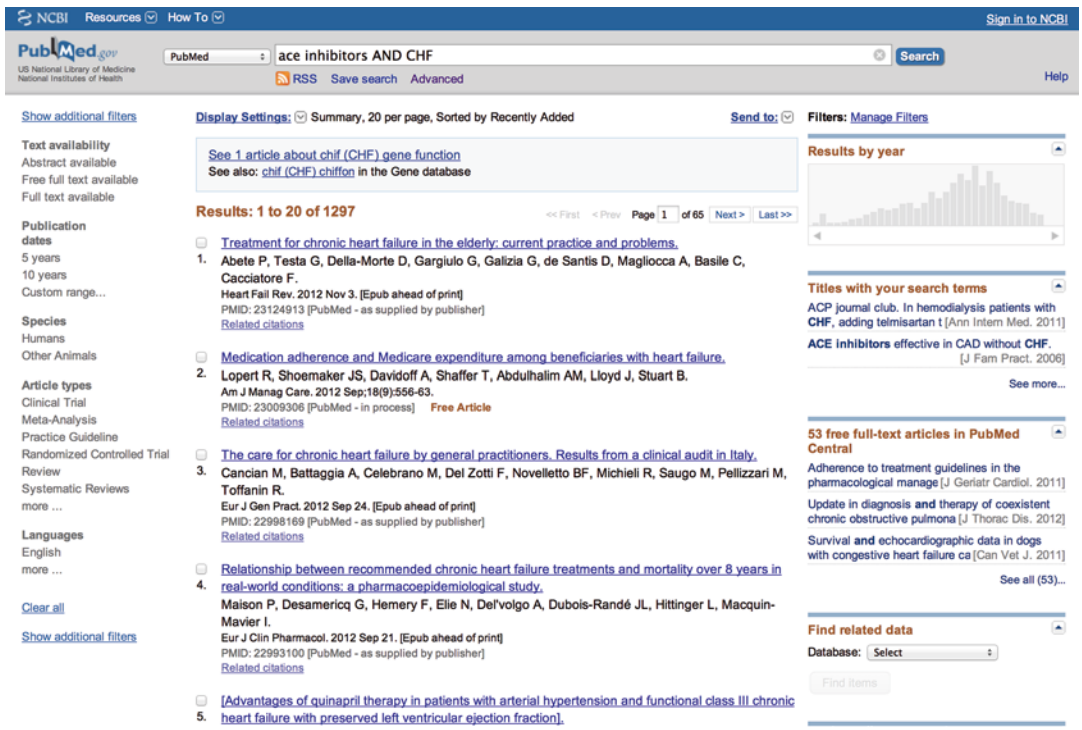

Fig. 21.4 Setting limits in PubMed. This screen shows some of the many limits that are available in PubMed, including article type, species type, subject subsets, availability of the article online, language of publication, gender, age group, and others (Courtesy of National Library of Medicine)

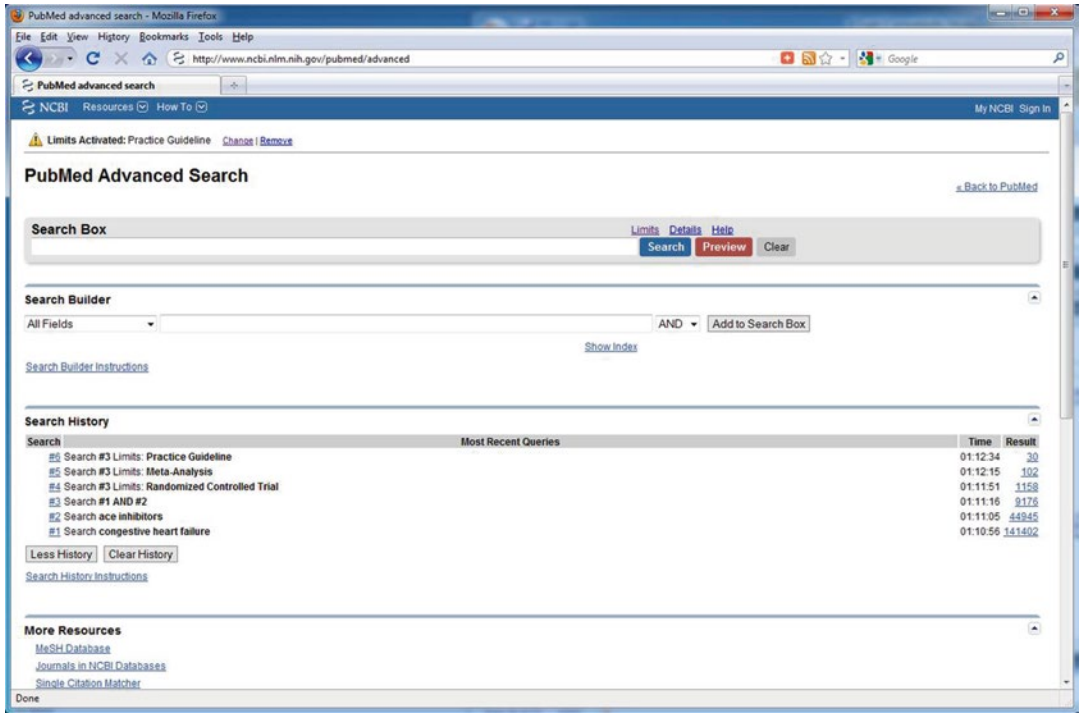

Fig. 21.5 Advanced search interface of PubMed, showing the use of sets and application of Boolean operators as well as limits. The user started with a search on the disease "congestive heart failure" and the drug class "ACE inhibitors." He or she subsequently narrowed the search by limited the output to "Randomized Controlled Trials," "Meta-Analysis," or "Practice Guidelines" (Courtesy of National Library of Medicine)

types, subsets, age ranges, and publication date ranges. These are accessed from the left-hand side of the results screen, with the most commonly used ones shown and the others accessible by additional mouse clicks.

As in most bibliographic systems, users can also search PubMed by building search sets and then combining them with Boolean operators to tailor the search. This is called the "advanced search" function of PubMed. Consider a user searching for studies assessing the reduction of mortality in patients with CHF through the use of ACE inhibitors. 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. Figure 21.5 shows the PubMed advanced search screen such a searcher might develop. This searcher has limited the output (using some of the limits listed in Fig. 21.4) with various publication types known to contain the best evidence for this question.

PubMed has another approach to finding the best evidence, which is through the use of its Clinical Queries function, 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 questions. These strategies arise from research assessing the ability of MEDLINE search statements to identify the best studies for therapy, diagnosis, harm, and prognosis (Haynes et al. 1994). The second approach to retrieving the best evidence aims to retrieve evidence-based resources that are syntheses and synopses, in particular meta-analyses, systematic reviews, and practice guidelines. The strategy derives in part from research by Boynton et al. (1998). 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.

As noted already, a great number of biomedical journals use the Highwire system for online access to their full text. The Highwire system provides a retrieval interface that searches over the complete online contents of 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 browse through specific issues as well as collected resources.

Once an article has been found, a wealth of additional features is available. First, the article is presented both in HTML and PDF form, with the latter providing a more readable and printable version. Links are also provided to related articles from the journal as well as the PubMed reference and its related articles. Also linked are all articles in the journal that cited this one, and the site can be configured to set up a notification e-mail when new articles cite the item selected. Finally, the Highwire software provides for "Rapid Responses," which are online letters to the editor. The online format allows a much larger number of responses than could be printed in the paper version of the journal. Other journal publishers use comparable approaches.

A growing number of search engines allow searching over many resources. The general search engines Google, Microsoft Bing, and others allow retrieval of any types of documents they have indexed via their Web crawling activities. Other search engines allow searching over aggregations of various sources, such as NLM Gateway,[58] which allows searching over all NLM databases and other resources in one simple interface.

21.6 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 searching in the systems they are using. Many frameworks have been developed to put the results in context. One of those frameworks organized evaluation around six questions that someone advocating the use of IR systems might ask (Hersh and Hickam 1998):

1. Was the system used?
2. For what was the system used?
3. How well did they use the system?
4. Were the users satisfied?
5. What factors were associated with successful or unsuccessful use of the system?
6. Did the system have an impact?

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

21.6.1 System-Oriented Evaluation

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. Recall is the proportion of relevant documents retrieved from the database:

$$\text{Recall} = \frac{\text{number of retrieved and relevant documents}}{\text{number of relevant documents in database}} \quad (21.5)$$

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 Equation (21.5) 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 be the

[58] http://gateway.nlm.nih.gov/

total number of unique, relevant documents identified by one or more searches on the query topic.

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

$$\text{Precision} = \frac{\text{number of retrieved and relevant documents}}{\text{number of documents retrieved}} \tag{21.6}$$

This measure answers the question, for a search, what fraction of the retrieved documents is 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 and McGill (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 0.0–1.0. 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. Instead, precision is measured at every point in the process in which a relevant document is obtained, and the MAP measure is found by averaging these points for the whole query.

A good deal of evaluation in IR is done via **challenge evaluations**, in which a common IR task is defined and a **test collection** of documents, topics, and **relevance judgments** are developed. The relevance judgments define which documents are relevant for each topic in the task, allowing different researchers to compare their systems with others on the same task and improve them. The longest running and best-known challenge evaluation in IR is the **Text REtrieval Conference** (**TREC**, trec.nist.gov), which is organized by the U.S. **National Institute for Standards and Technology** (**NIST**,).[59] 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(s). TREC is organized into tracks geared to specific interests. A book summarizing the first decade of TREC provides more information on this important IR initiative that is still ongoing (Voorhees and Harman 2005).

While TREC has been focused on general IR, there was a track that ran for several years devoted to retrieval from genomics resources (Hersh et al. 2006; Hersh and Voorhees 2009). In addition, more recently TREC has added a track focused on medical records based on the use case of identifying patients as potential candidates for clinical studies (Voorhees and Tong 2011; Voorhees and Hersh 2012).

Some researchers have criticized or noted the limitations of relevance-based measures. 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 (Harter 1992; Swanson 1988). Hersh (1994) has noted

[59] www.nist.gov

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).

21.6.2 User-Oriented Evaluation

What alternatives to relevance-based measures can be used for determining performance of individual searches? Harter admits that if measures using a more situational view of relevance cannot be developed for assessing user interaction, then recall and precision may be the only alternatives. Some alternatives have focused on users being able to perform various information tasks with IR systems, such as finding answers to questions (Egan et al. 1989; Hersh and Hickam 1995; Hersh et al. 1996; Mynatt et al. 1992; Wildemuth et al. 1995). For several years, TREC featured an Interactive Track that had participants carry out user experiments with the same documents and queries (Hersh 2001). 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. (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 showed that the experienced clinicians and librarians achieved comparable recall in the range of 50 %, although the librarians had 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.

A follow-up study yielded some additional insights about the searchers (McKibbon 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.

Recognizing the limitations of recall and precision for evaluating clinical users of IR systems, Hersh and coworkers 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.

The first study by this group using the task-oriented approach compared Boolean versus natural language searching in the textbook Scientific American Medicine (Hersh and Hickam 1995). Thirteen medical students were asked to answer ten 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 to 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) and Knowledge Finder (KF; Hersh et al. 1996). These systems represented the ends of the spectrum in terms of using Boolean searching on human-indexed thesaurus terms (Ovid) versus natural language searching on words in the title, abstract, and indexing terms (KF). 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 % correctly before searching and 85.4 % correctly 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 systems can be used equally well with minimal training.

A more 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 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 attempted to measure what factors might influence searching. A multitude of factors, such as age, gender, computer experience, and time taken to search, were not associated with successful answering of questions. Successful answering was, however, 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 a complete lack of association with ability to answer these questions.

Two studies have extended this approach in various ways. Westbook et al. (2005) assessed use of an online evidence systems and found that physicians answered 37 % of questions correctly before use of the system and 50 % afterwards, while nurse specialists answered 18 % of questions correctly and also 50 % afterwards. Those who had correct answers before searching had higher confidence in their answers, but those not initially knowing the answer had no difference in confidence whether their answer turned out to be right or wrong. McKibbon and Fridsma (2006) performed a comparable study of allowing physicians to seek answers to questions with resources they normally use employing the same questions as Hersh et al. (2002). This study found no difference in answer correctness before or after using the search system. Clearly these studies show a variety of effects with different IR systems, tasks, and users.

Pluye and Grad (2004) performed a qualitative study assessing impact of IR systems on physician practice. The study identified four themes mentioned by physicians:

- Recall—of forgotten knowledge
- Learning—new knowledge
- Confirmation—of existing knowledge
- Frustration—that system use not successful

The researchers also noted two additional themes:

- Reassurance—that system is available
- Practice improvement—of patient-physician relationship

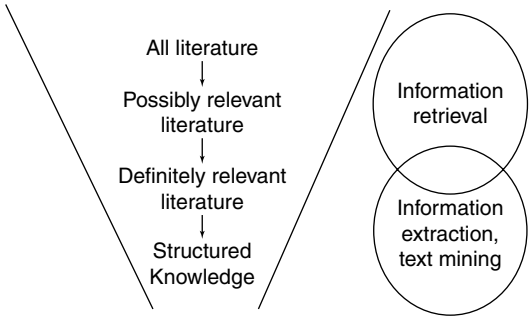

Fig. 21.6 Funnel of knowledge discovery, showing how an information need starts with a search (information retrieval) leading to a large possibly relevant set of literature that is winnowed down to a smaller definitely relevant set (usually by human inspection but with techniques like information extraction and text mining possibly automating the process in the future). Ultimately actionable knowledge is obtained that can be applied by a human or fashioned into, for example, rules for a computer-based decision support system (Reproduced with permission of Springer (Hersh 2009))

21.7 Research Directions

The above evaluation research shows that there is still plenty of room for IR systems to improve their abilities. In addition, there will be new challenges that arise from growing amounts of information, new devices, and other new technologies.

There are also other areas related to IR where research is ongoing in the larger quest to help all involved in biomedicine and health—including patients, clinicians and researchers—to better apply knowledge to improve health. Figure 21.6 shows this author's "funnel" by which the user searches all of the scientific literature using IR systems to obtain a set of possibly relevant literature. In the current state of the art, he/she reviews this literature by hand, selecting which articles are definitely relevant and may become "actionable knowledge" that can be acted upon to make better decisions.

Our ability to carry out the activities in the upper part of the funnel, i.e., IR, is much better than those in the lower part. These areas include:

- **Information extraction a**nd **text mining**—usually through the use of natural language processing (NLP, see Chap. 8) to extract facts and knowledge from text. These techniques are often employed to extract information from the EHR, with a wide variety of accuracy as shown in a recent systematic review (Stanfill et al. 2010).
- **Summarization**—Providing automated extracts or abstracts summarizing the content of longer documents (Fiszman et al. 2004)
- **Question-answering**—Going beyond retrieval of documents to providing actual answers to questions, as exemplified by IBM's Watson system (Ferrucci et al. 2010).

21.8 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 (1999) noted that libraries of both types elicited different definitions of what they actually are, with researchers tending to view libraries as content collected for specific communities and librarians alternatively viewing them as institutions or services. Lindberg and Humphreys (2005) laid out a vision in 2005 for libraries 10 years hence, noting that while collections would be virtual and accessed in many diverse ways, other elements of science would stay intact, including journals and the peer review process.

This section provides an overview of key issues of digital libraries, with an orientation toward biomedical libraries.

21.8.1 Functions and Definitions of Libraries

The central function of libraries is to maintain collections of published literature. They may also store unpublished literature in archives, such as

letters, notes, and other documents. The general focus on published literature has implications. One of these is that, for the most part, quality control can be taken for granted. Until recently, most published literature came from commercial publishers and specialty societies that had processes such as peer review, which, although imperfect, allowed the library to devote minimal resources to assessing their quality. While libraries can still cede the judgment of quality to these information providers in the Internet era, they cannot ignore the myriad of information published only on the Internet, for which the quality cannot be presumed.

Other functions of libraries besides maintaining collections include cataloging and classification of items in those collections, being a place (even virtual) where individuals could go to get assistance with information seeking, and providing space for work or study, particularly in universities.

The paper-based nature of traditional libraries carrieda number of assumptions that are challenged in the digital era. For example, items were produced in multiple copies, freeing the individual library from excessive worry that an item could not be replaced. In addition, items were fairly static, simplifying their cataloging. With digital libraries, this status quo is challenged. There is a great deal of concern about archiving of content and managing its change when fewer "copies" of it exist on the file servers of publishers and other organizations. A related problem for digital libraries is that they do not own the "artifact" of the paper journal, book, or other item. This is exacerbated by the fact that when a subscription to an electronic journal is terminated, access to the entire journal is lost; that is, the subscriber does not retain accumulated back issues, as was taken for granted with paper journals.

21.8.2 Access

Probably every Web user is familiar with clicking on a Web link and receiving an error message that a page cannot 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 hyper-link, linked to a **uniform resource name** (**URN**) that would be persistent (Sollins and Masinter 1994). The combination of a URN and a URL, a **uniform resource identifier** (**URI**), would provide persistent access to digital objects. However, no publicly available resource for resolving URNs and URIs was ever implemented on a large scale.

One approach that has seen widespread adoption by publishers, especially scientific journal publishers, is the **digital object identifier** (**DOI**,)[60] (Paskin 2006). The DOI has recently been given the status of a standard by the NISO with the designation Z39.84. The DOI itself is relatively simple, consisting of a prefix that is assigned by the International DOI Foundation (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. Publishers are encouraged to facilitate resolution by encoding the DOI into their URLs in a standard way, e.g., http://dx.doi.org/10.1197/jamia.M0996 for a paper cited earlier in the chapter (Hersh et al. 2002).

21.8.3 Interoperability

As noted throughout this chapter, metadata is a key component for accessing content in IR systems. It takes on an 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

[60] www.doi.org

et al. note that three levels of agreement must be achieved in digital libraries:

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

21.8.4 Intellectual Property

Intellectual property issues are a major concern in digital libraries. Intellectual property is difficult to protect in the digital environment because although the cost of production is not insubstantial, the cost of replication is near nothing. Furthermore, in circumstances such as academic publishing, the desire for protection is situational. For example, individual researchers may want the widest dissemination of their research papers, but each one may want to protect revenues realized from synthesis works or educational products that are developed. The global reach of the Internet has required that intellectual property issues be considered on a global scale. The **World Intellectual Property Organization** (**WIPO**,)[61] is an agency of the United Nations devoted to developing worldwide policies, although understandably, there is considerable diversity about what such policies should be.

21.8.5 Preservation

Another function of libraries of all types is preservation of materials. In paper-based libraries, the goal of preservation was the survival of the physical object, i.e., the book, journal, image, etc. that could become lost, stolen, or deteriorated. Preservation issues in digital libraries are somewhat different. Digital libraries still do need to be concerned with physical survival of the information. Lesk (2005) compared the longevity of digital materials. He noted that the longevity for magnetic materials was the least, with the expected lifetime of magnetic tape being 5–10 years. Optical storage has somewhat better longevity, with an expected lifetime of 30–100 years depending on the specific type. Ironically, paper has a life expectancy well beyond all these digital media. Rothenberg (1999) has noted that the Rosetta Stone, which provided help in interpreting ancient Egyptian hieroglyphics and has survived over 20 centuries. He reiterated Lesk's description of the reduced lifetime of digital media in comparison with traditional media, and to note another problem familiar to most long-time users of computers, namely, data can become obsolete not only owing to the medium, but also as a result of data format. Both authors noted that storage devices as well as computer applications, such as word processors, have seen their formats change significantly over the last couple of decades.

The US Library of Congress has devoted considerable effort to digital preservation, documenting its efforts on the Web site.[62] The largest preservation effort in the US is National Digital Information Infrastructure Preservation Program (NDIIPP)[63] of the Library of Congress. Other digital preservation efforts include Portico,[64] a collaboration of publishers, libraries, and government agencies to preserve electronic scholarly content and LOCKSS (Lots of Copies Keep Stuff Safe,[65] which provides libraries with digital preservation tools and support. An effort related to the latter is CLOCKSS,[66] which is a "trusted community-governed archive."

21.9 Future Directions for IR Systems and Digital Libraries

There is no doubt that considerable progress has been made in IR and digital libraries. Seeking online information is now done routinely not

[61] www.wipo.org

[62] www.digitalpreservation.gov

[63] www.digitalpreservation.gov

[64] www.portico.org

[65] www.lockss.org

[66] www.clockss.org

only by clinicians and researchers, but also by patients and consumers. There are still considerable challenges to make this activity more fruitful to users. They include:

- How do we lower the effort it takes for clinicians to get to the information they need rapidly in the busy clinical setting?
- How can researchers extract new knowledge from the vast quantity that is available to them?
- How can consumers and patients find high-quality information that is appropriate to their understanding of health and disease?
- Can the value added by the publishing process be protected and remunerated while making information more available?
- How can the indexing process become more accurate and efficient?
- Can retrieval interfaces be made simpler without giving up flexibility and power?
- Can we develop standards for digital libraries that will facilitate interoperability but maintain ease of use, protection of intellectual property, and long-term preservation of the archive of science?

Suggested Readings

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern information retrieval: The concepts and technology behind search* (2nd ed.). Reading: Addison-Wesley. A book surveying most of the automated approaches to information retrieval.

Croft, W., Metzler, D., & Strohman, T. (2009). *Search engines: Information retrieval in practice*. Boston: Addison-Wesley. A book surveying most of the automated approaches to search engines.

Frakes, W., & Baeza-Yates, R. (Eds.). (1992). *Information retrieval: Data structures and algorithms*. Englewood Cliffs: Prentice-Hall. 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. (2009). *Information retrieval: A health and biomedical perspective* (3rd ed.). New York: Springer. A textbook on information retrieval systems in the health and biomedical domain that covers state-of-the-art as well as research systems.

Lindberg, D., & Humphreys, B. (2005). 2015 – The future of medical libraries. *New England Journal of Medicine, 352*, 1067–1070. A vision of the future of medical libraries from two leaders of the NLM.

Miles, W. (1982). *A history of the National Library of Medicine: The nation's treasury of medical knowledge*. Bethesda: U.S. Department of Health and Human Services. 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.

Straus, S. E., Glasziou, P., et al. (2010). *Evidence-Based Medicine: How to Practice and Teach it* (4th ed.). New York, NY, Churchill Livingstone.

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 you think open-access publishing will succeed or not.
3. How would you aggregate the clinical evidence-based resources described in the chapter into the best digital library for clinicians?
4. Devise a curriculum for teaching clinicians and patients the most important points about searching for health-related information.
5. Find a consumer-oriented Web page and determine the quality of the information on it.
6. What are the limitations of recall and precision as evaluation measures and what alternatives would improve upon them?
7. Select a concept that appears in two or more clinical terminologies and demonstrate how it would be combined into a record in the UMLS Metathesaurus.
8. Describe how you might devise a system that achieves a happy medium between of intellectual property and barrier-free access to the archive of science.