Chapter 9
INFORMATION RETRIEVAL AND DIGITAL LIBRARIES

William R. Hersh

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

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

Keywords
Information retrieval; digital library; indexing; controlled vocabulary; searching; knowledge-based information
1. OVERVIEW OF FIELDS

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

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

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

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