LECTURE TRANSCRIPT

INFORMATION RETRIEVAL: ACCESS TO KNOWLEDGE-BASED RESOURCES

William Hersh, M.D.

SLIDE 1

Welcome to the Health Informatics Building Block Information Retrieval Access To Knowledge-based Resources. I'm Dr. William Hersh from Oregon Health and Science University.

SLIDE 2

In this Health Informatics Building Block, we will begin with definitions of the field of information retrieval. We will then look at the components of information retrieval systems as well as discuss types and examples of knowledge-based resources. These include bibliographic and full-text, annotated and aggregated resources.

SLIDE 3

Information retrieval is the field concerned with the organization and retrieval of knowledge-based information. Information retrieval focuses mainly on textual information, but with the growing amount of multimedia such as images, sounds and video as well as more complex databases, the nature of IR has changed. Historically information retrieval has not been focused on patient-based information, but this is changing as well, particularly as information about patients is used to connect to knowledge-based information. IR is also sometimes called “search” and is probably the most prevalent activity on the World Wide Web by both clinicians and patients alike.

SLIDE 4

This slide shows the components of IR systems. Probably the most important of all these boxes is content. This is what the user seeks to retrieve. That content is usually retrieved by queries that the user enters, statements of information needs. The content in queries are matched by metadata which describes the content, as well as the user need, and a search engine connects the two. There are two intellectual processes associated with IR -- indexing and retrieval.

SLIDE 5

Indexing is the process of assigning metadata to content items. Most commonly metadata consists of subjects or terms. These are words or phrases that describe the content. The metadata might be individual words that occur in the content or they may be phrases that come from a controlled vocabulary and have been assigned by human indexers. Other metadata is assigned from content items and this can include things like the name of the author the source of the information, the publication type such as whether it's a journal article, or what type of study it is such, is a randomized controlled trial retrieval. Then there is the process of entering a query in the metadata and having the search engine match the content items. There are two common approaches to retrieval: there is Boolean searching with the use of the operators “and,” “or,” and “not” that give the user much more control over their retrieval. But more frequently, systems use natural language searching which sometimes applies Boolean operators in the background, but in essence the natural language approach takes words that the user types into the query and matches them to words in the content.

SLIDE 6

Another way to view IR is as part of knowledge discovery. That is when an individual has an information need, all of the literature goes into a funnel. Usually through the help of an information retrieval system, an individual will do a search and have a collection of possibly relevant literature. Then, usually by hand, the searcher will go through the retrieve literature and find that which is definitely relevant to his or her information need. Often times the process will stop at that point and he or she will use the information that's been retrieved for their decision-making process. Increasingly, however, we want to structure knowledge. We may wish to use in a clinical decision support system that's part of an electronic health record or we may want to automate the discovery of knowledge by processing the text of literature. The the upper part of the funnel is the process of information retrieval whereas the lower part of the funnel is the process of information extraction or which is sometimes called “text mining.” The bottom part of the funnel at this point in time is much more experimental. That is why the use of information retrieval systems is common and takes place every day, the use of information extraction and text mining is still a research process.

SLIDE 7

We will spend the remainder of this Health Informatics Building Block talking about knowledge-based resources and the systems that allow their retrieval. Before we can do that, however, it's helpful to classify the different types of resources. Probably the original and still most common type of knowledge-based resources is bibliographic content. **Bibliographic** content is usually rich in metadata that facilitate its retrieval. Bibliographic content, however, often only provides a description of the resource and not the actual resource itself. For that reason we have a growing interest in the **full-text** of knowledge-based resources, things like textbooks and journals and websites. Some knowledge-based resources are not merely textual, things like images and videos, and for this reason we need to think about **annotated**

knowledge-based resources where the annotation describes what is in the content. Finally there are **aggregations** that bring all of these different resources together and make them useful as a larger collection.

SLIDE 8

Let's begin by looking at **bibliographic** content. The original type of bibliographic content and still probably the most prevalent are bibliographic databases. Bibliographic databases have evolved over time. The old tried-and-true bibliographic databases such as MEDLINE have been revitalized with new features. In addition, newer types of bibliographic databases have emerged such as the National Guidelines Clearinghouse. Another type of bibliographic content is Web catalogs. These are collections of Web resources that share many characteristics of traditional bibliographic databases. Finally a new approach to bibliographic content is known as RSS which stands for alternatively Real Simple Syndication or Rich Site Summary. RSS is usually provided in feeds that provide information about new content.

SLIDE 9

An important point to remember about bibliographic databases is that they usually consist of metadata and may not consist of the resource itself even though in this day and age the resource itself may be one click away from the bibliographic information. In the old days many of you might remember the system might only have bibliographic information or even in the olden days the bibliographic information might be in a book or card catalog in a library. A great deal of bibliographic databases are produced by the US government. Databases like MEDLINE, AIDSLINE, Cancerlit and Toxilit, but bibliographic databases are also produced by commercial publishers such as CINAHL, EMBASE, and Current Contents.

SLIDE 10

Certainly the granddaddy of all bibliographic databases as MEDLINE. MEDLINE is most commonly accessed through a software system at the National Library of Medicine called PubMed, but can be accessed other ways. MEDLINE contains references to biomedical journal literature. It really was the original medical information retrieval application. Since 1998, MEDLINE has been freely available to the entire world by the PubMed system, accessible at pubmed.gov. MEDLINE is one of the many bibliographic databases produced by the US National Library of Medicine. It contains over 19 million references to peer-reviewed literature dating back to 1966 and even some additional literature before then. It covers about 5000 journals, mostly, but not exclusively in English language and roughly has about 600,000 new references added yearly. Modern MEDLINE, of course, has links to the full-text of articles and other resources that might be associated with the bibliographic record, but clearly MEDLINE itself is a bibliographic database.

SLIDE 11

A newer bibliographic database is the National Guidelines Clearinghouse. This is produced by the Agency for Healthcare Research and Quality and is available at www.guideline.gov. This database contains detailed information about clinical practice guidelines. For example, it includes the degree to which they are evidence-based and also provides an interface that allows comparison of elements in the database for more than one guideline. The National Guidelines Clearinghouse has links to those that are freely available on the web or links to the producers of the guideline when the guideline is proprietary and not necessarily freely available.

SLIDE 12

Another type of bibliographic resources is **web catalogs**. These generally aim to provide quality filtered websites aimed at specific audiences. Some, for example, are aimed towards clinicians such as HON Select and Translating Research into Practice. Other web catalogs are aimed more towards patients or consumers such as HealthFinder.

SLIDE 13

The final type of bibliographic content we will discuss is RSS. Again RSS feeds provide the short summaries which may consist of news or recent postings or medical journal articles, any other type of information and for which the RSS feed is a structured summary. Users then receive these feeds by ane aggregator which can be either a stand alone application or, increasingly, e-mail clients will accept RSS feeds so the user can configure them for which sites they want to receive and even filter them based on the content that's in the feed. There are two different versions of RSS, but each basically provide the following information:

- there's a title which is the name of the item;
- a link which is the URL that provides a pointer to where the full information resource is;
- and, then a description that is a brief description of that full information resource.

SLIDE 14

The second major type of content we see in information retrieval systems is **full-text content**. This of course contains the complete text of the resource as well as any tables, figures, images, or other content. If there is a corresponding print version of a full-text resource both are usually identical. So for example, most medical journals now will provide the full-text of content actually not only as a regular HTML page but will also provide something like a PDF file that has an exact replica of what the printed journal might look like. Full-text content includes periodicals, so journals, newspapers, magazines etc. Many, many books are now available in their full-text form and of course there are many websites that have mainly textual information to them and they are included in this category as well.

SLIDE 15

Certainly one of the major types of full-text online content of interest to medical personnel is the full-text of the medical literature. Almost all biomedical journals are now available electronically many of them are published by High Wire Press which adds value to the content of the original publisher and includes such well-known journals as British Medical Journal, Journal of the American Medical Association, New England Journal of Medicine and others. There are a growing number of journals that are available via the open access model where the content is freely available. These include the families of journals from BioMed Central and Public Library of Science. Some publishers will license full-text content and provide it to vendors who make collections of content available. A couple of well-known companies that do this are Ovid which is a core collection of 60 to 80 major journals and then many more as well as MDConsult which has many journals, but mostly less prestigious ones. Interestingly the impediments to wider dissemination of primary literature are no longer technical. We know how to publish electronic medical journals but there continue to be economic challenges such as how you pay for the production of content and recoup those costs through various kinds of economic models.

SLIDE 16

Another type of full-text content is books. Most of the well-known clinical textbooks are now available electronically. Such tried-and-true volumes like *Harrison's Principles Of Internal Medicine.* The National Library of Medicine has developed the book site that is actually part of the larger PubMed system. There are also many compendia of drugs, diseases, evidence, and so forth that are available and of course the handbooks that historically have been very popular for clinicians to carry in the pockets of white coats.

SLIDE 17

It's important to remember that online versions of periodicals and books can have a great deal of value added. For example multimedia can demonstrate clinical findings that can't just be described in textbooks. So for example the different types of skin lesions, the shuffling gait of a patient with Parkinson's disease, and so forth. Of course electronic books can be bundled so that multiple books are in the collection and searching can take place over many of them. Electronic books also can be updated more frequently than just with the new edition that comes out of the printing press. And they can also provide linkage to other information such as references, self assessments, updates, and other kinds of resources.

SLIDE 18

The final type of full-text contents will discusses **websites**. And when we say websites here, we're not talking about all websites, but a narrower subset of websites that contain coherent collections of information. These two of course usually take advantage of web features such as linking and multimedia.

SLIDE 19

In this and several following slides will talk about some notable full-text content that's available on websites. One of the leading producers of health-related content on the web are US government agencies. So for example CancerNet from the National Cancer Institute provides information about cancer, the Centers for Disease Control has a substantial variety of information available related to travel, infection, and other health related issues. Many of the institutes of the National Institutes of Health or NIH, for example, the National Heart Lung And Blood Institute have very well-developed websites with a great deal of content.

SLIDE 20

There are many other websites from either commercial or nonprofit entities. There are a number of websites oriented to physicians and other healthcare professionals providing news and overviews these include examples such as Medscape and PEPID. Also, many professional societies provide health-related information to their members. There are also websites that are more patient or consumer oriented. Of note are Intelihealth and NetWellness.

SLIDE 21

There are many more types of web content as well. One well-known source is Wikipedia which is an encyclopedia that has free access to everyone and is sustained by a worldwide distributed authorship. There have been concerns about manipulation of the facts in Wikipedia. However over time the quality of information has been found to be as good as traditional encyclopedias. With health-related information the quality has found been found to be reasonably good. And it's important that it is good because Wikipedia pages often appear at the top of many web search engines. Another type of web content is the body of knowledge and this is best exemplified by the software engineering body of knowledge which organizes knowledge about the field. One other type of web content of note is weblogs or “blogs.” These are ongoing web-based commentaries on many topics. They vary in quality depending on the interests and motivation of the author of the blog and they demonstrate the ability of the web to amplify information or also, unfortunately, to amplify misinformation.

SLIDE 22

The third major type of knowledge-based resources is **annotated** resources. These typically consist of either non-text or structured text information that is then annotated with text that describes what that information is about. The types of annotated resources include image collections, citation databases, evidence-based medicine databases, genomics databases, and other databases. There are many image collections available on the web. Naturally they are most prominent in the visual medical specialties such as radiology, pathology and dermatology. Some of the better-known collections include the Visible Human which was developed by the National Library of Medicine and contains both photographic and radiologic slices cross-section of a man and a woman. BrighamRad is a well-known radiology image

collection likewise WebPath and PIER, are well-known hepatology image collections and DermIS is a well-known collection of skin related problems. Many of the images in these collections have associated text and of course that assists with their indexing and retrieval making it possible to find the image that one is searching for.

SLIDE 23

There are also **citation databases**. These are databases of articles that cite other articles. The original citation database was the Science Citation Index which was followed by the Social Science Citation Index. These are both databases of journal articles that are cited by other journal articles. These are actually now part of a package called Web of Science that itself is part of a larger project called Web of Knowledge. Other citation databases are available. There is one called SCOPUS and the Google Scholar system provides a form of citation databases in that there are linkages between the references in one article and the actual papers that are in other articles.

SLIDE 24

There are many evidence-based medicine databases. These are typically collections of information that has gone through some process where the tenets of evidence-based medicine are applied. Probably the best-known of these is the Cochrane Database of Systematic Reviews which is a larger and ever increasing collection of systematic reviews done on different clinical topics mainly focused on medical interventions. It is a requirement for the authors of the systematic reviews to keep them up-to-date. The BMJ maintains a product called Clinical Evidence which it describes as an evidence formulary so it has more concise summaries of evidence oftentimes coming from Cochrane database reviews. Up to Date could somewhat fall into the full-text textbook category, but it does to apply the principles of evidence-based medicine to give clinically oriented overviews of different medical topics. The Physicians Information and Education Resource or PIER is a resource from the American College of Physicians that has disease oriented overviews and each individual statement in the collection is tagged for its level of evidence. Finally there's a resource called InfoPOEMS that provides as it describes itself Patient Oriented Evidence that Matters.

SLIDE 25

As we continue to sequence genomes and collect information related to the function of genomes, we see the continued growth of genomics databases. Clearly the leading organization for genomic information is the National Center for Biotechnology Information or NCBI. NCBI is the unit within the National Library of Medicine that develop PubMed so the NCBI collections include literature references. However there are also many other genomics related information resources. There is a textbook of genetic diseases call the On-Line Mendelian Inheritance in Man or (OMIM). There is GenBank which is a database of gene sequences that have been discovered experimentally. There are various databases that show the structure of DNA or proteins. This is called the Molecular Modeling Database. There are genomes which

provide catalogs of genes and maps that show the locations of genes on chromosomes in many different organisms.

SLIDE 26

Some other important annotated databases include ClinicalTrials.gov which is a product of the National Library of Medicine and was originally a database of clinical trials that were funded by NIH with the goal being that individuals or their clinicians could find trials for patients to enroll in. However the use of ClinicalTrials.gov has expanded and now serves as a register of all clinical trials and in fact medical journals will not publish clinical trials if they are not prospectively registered in clinicaltrials.gov. This prevents the investigators carrying out clinical trials from manipulating them such as not reporting certain results. Another annotated databases is NIH Reporter which is a database of all research grants funded by NIH and it replaced the previous system that provided this function called CRISP.

SLIDE 27

A final type of knowledge-based information resource used in information retrieval systems is **aggregations.** These typically integrate many resources from the three types that we've already described. On this slide I will talk about three audiences for aggregations and an example of an aggregation for that audience. For clinicians there is something like Merck Medicus which provides a collection of many resources that are made available to any licensed physician in the United States. Biomedical researchers tend to be focused more on the area where they do their research. One of the most common type of aggregations for this audience is the Model Organism Database. This is a database of all the information, typically genomics, articles, images and so forth for the particular organism that study. One of the best-known model organism databases is the Mouse Genome Informatics Database which contains information related to the genome of the mouse. There are also aggregations for consumers. Probably the best-known consumer aggregation is MEDLINEplus which is maintained by the National Library of Medicine and it integrates a variety of license resources and public websites for healthcare consumers.

SLIDE 28

Thank you for listening to this Health Informatics Building Block. This work is provided under the terms of a Creative Commons Public license. The work is protected by copyright and/or other applicable law. Any use of the work other than as authorized under this license or copyright law is prohibited. We hope you will provide feedback on this Health Informatics Building Block so we can improve it. Thank you.