Information Retrieval: Grand Challenges in the 21st Century

William Hersh, M.D.
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
hersh@ohsu.edu
www.billhersh.info

References

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

Hersh, W. (1999). "A world of knowledge at your fingertips": the promise, reality, and future directions of on-line information retrieval. *Academic Medicine*, 74: 240-243.

Haynes, RB (2001). Of studies, syntheses, synopses, and systems: the "4S" evolution of services for finding current best evidence. *ACP Journal Club*, 134: A11-A13.

Gorman, P. (1995). Information needs of physicians. *Journal of the American Society for Information Science*, 46: 729-736.

Bero, L. and Rennie, D. (1996). The Cochrane Collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Journal of the American Medical Association*, 274: 1935-1938.

Anonymous (1990). *IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries*. Piscataway, NJ. IEEE Press.

Lagoze, C. and Van de Sompel, H. (2001). The Open Archives Initiative: building a low-barrier interoperability framework. *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*, Roanoke, VA. ACM Press. 54-62.

Hersh, W. and Rindfleisch, T. (2000). Electronic publishing of scholarly communication in the biomedical sciences. *Journal of the American Medical Informatics Association,* 7: 324-325.

Weiss, R. (2003). A fight for free access to medical research. Washington Post. August 5, 2003. A01.

Healy, B. (2003). Power to the people! US News & World Report. September 8, 2003. 43.

McLellan, F. (2003). US bill says government funded work must be open access. *Lancet,* 362: 52.

Hersh, W. (2001). The way of the future? Review of Biomed Central. *Nature,* 413: 680.

Butler, D. (2003). Who will pay for open access? *Nature,* 425: 554-555.

Martin, M., Kuhlman, D., et al. (2002). Federated digital rights management: a proposed DRM solution for research and education. *D-Lib Magazine,* 8: 7. http://www.dlib.org/dlib/july02/martin/07martin.html.

Information Retrieval: Grand Challenges in the 21st Century

William Hersh, M.D.
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
hersh@ohsu.edu
www.billhersh.info

Overview

- Quick primer on information retrieval
- Non-grand challenges
- Grand challenges
- Final thoughts

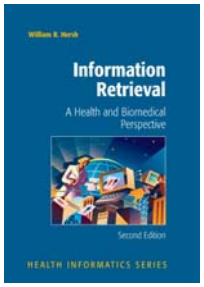
Quick primer on information retrieval

Overview of clinical information

- Two basic types, with different uses and applications
  - *Patient-specific* information is generated in the care of patients
    - Applications: electronic health records, telemedicine, etc.
  - *Knowledge-based* information is the scientific literature of health care
    - Applications: information retrieval systems, evidence-based medicine

Information retrieval

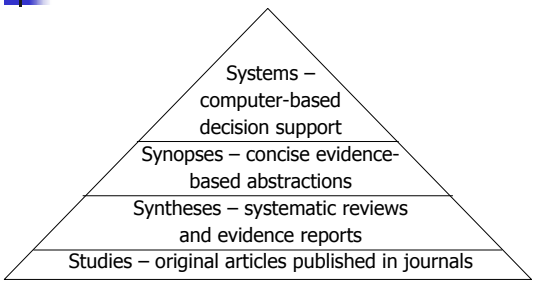
- Focuses on indexing and retrieval of knowledge-based information
- Historically centered on text in documents, but increasingly associated with multimedia and even patient-specific information
- www.irbook.info



A good organizing principle: evidence-based medicine

- Evidence-based medicine (EBM) is the application of best scientific evidence to clinical decision making
- Focus has changed in recent years (Hersh, 1999)
  - First generation EBM focused on de novo retrieval and appraisal of individual studies
  - Next generation EBM aims to develop syntheses and synopses for users

Another way to view EBM (Haynes, 2001)



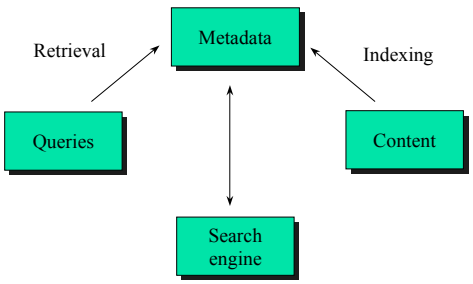
Searching – everyone is doing it …

… everyone knows about it …

(Am I a lucky father or what?)

… but new problems have emerged

IR system



The intellectual tasks of IR

- Indexing
  - Assigning metadata to content items
  - Can assign
    - Terms – words, phrases from controlled vocabulary
    - Attributes – e.g., author, source, publication type
- Retrieval
  - Most common approaches are
    - Boolean – use of AND, OR, NOT
    - Natural language – words common to query and content

Non-grand challenges

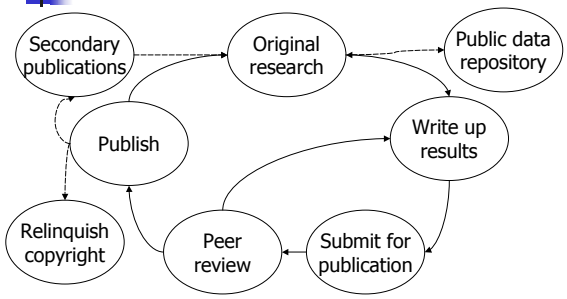
- While occasionally challenging, these problems are, for the most, not grand challenges
  - Finding a known item (Google Toolbar anyone?)
  - Using an IR system to search a single collection of content (e.g., local Web site, textbook, maybe even MEDLINE)
- These do not obviate need for users better learning how search systems work, but are not major research challenges
  - Major role for librarians and informaticians

Grand challenges for Information Retrieval

- Covered in this talk
  - Content – the right information for the right task
  - Linkage – across multiple resources
  - Access – open access but protective of intellectual property
- Others of great interest, for another day
  - Indexing – metadata for Web content
  - Evaluation – best measures, meaningful studies
  - Quality of health information on the Web

Content

The life-cycle of knowledge-based information



A classification of knowledge-based content (Hersh, 2003)

- Bibliographic
  - By definition rich in metadata
- Full-text
  - What users want
- Databases/Collections
  - Specialized content
- Aggregations
  - The power of the Web

Most important issues concerning content

- The right information for the right task
- Consider the clinician, who wants
  - Synopsis at the point of care
  - Synthesis as an entry point to explore questions further
  - Original studies when want to drill down to basic assumptions
  - Overviews and "background" information to refresh knowledge or explore new areas

How do we produce refined knowledge?

- Understanding users' questions, e.g., Gorman, 1995 and others
- Mass production is challenging
  - Academic rewards not clear
  - Pure volunteerism, e.g., Cochrane Collaboration (Bero, 1996), is too slow
  - Business model challenging as well, i.e., need linkage with other resources…

Linkage

Consider this scenario

- A primary care clinician of an elderly patient who has hypertension, congestive heart failure, sleep apnea, and obesity
  - Has charted pertinent information in electronic health record
  - Now wants recommendations from a guideline with overview of supporting evidence
  - Later wants to explore recommendations in more detail, including reading systematic review and some original clinical trials it has included
  - May want basic review of topics seen infrequently in practice

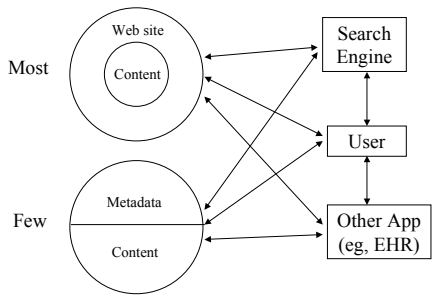
Some impediments for this clinician

- Cannot link directly from guideline to supporting or background information
- Wants to access pertinent section of a favorite textbook directly
  - Does not want to go to each Web site, log on, and use site search engine
- Would like to navigate across levels of evidence from compendium to systematic review to original clinical trial or other study
- May want to create personal digital library of preferred content

Impediments for others

- Publishers
  - Might desire to allow access to pieces of content but need assurances of revenue and intellectual property protection
- Content aggregators
  - Want to "mix and match" content that is "best of breed" but difficult to do across content of different publishers

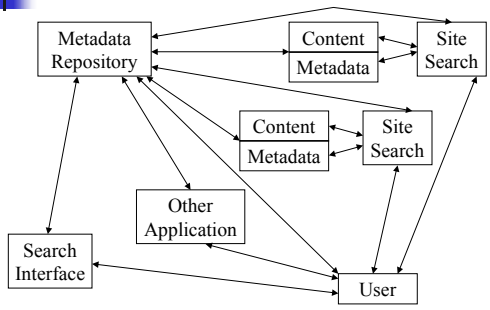
The current problem: most information is in *silos*



Overcoming the impediments: Interoperability

- IEEE, 1990: "Ability of two or more systems…to exchange information or use the information that has been exchanged"
- Used in digital library community to describe seamless access and integration
- Required to facilitate IR interoperability are
  - Minimum set of metadata and interapplication interfaces
  - Cooperation among publishers, vendors, and others to agree upon standards

From silos to interoperability



How might we achieve this?

- A starting point is Open Archives Initiative (OAI, www.openarchives.org)
- OAI promotes the "exposure" of archives' metadata such that systems can know what content is available and how it can be *harvested* (Lagoze, 2001)
- Each record in an OAI collection contains metadata
  - Protocol has "verbs" for metadata harvesting
  - Example: http://www.purl.org/NET/oai_explorer

Are there any good examples of integrated resources?

- Yes, from the genomics community
- Databases of National Center for Biotechnology Information (NCBI) are linked
  - Literature
  - Nucleotide and protein sequences
  - Protein structures
  - Textbooks and other textual resources
  - Genomes and map
- Which leads us to issues of access…

Access

Many issues, but we will focus on electronic publishing

- Impediments to wider dissemination are economic and political, not technical (Hersh, 2000)
  - Journals have monopolies due to promotion and tenure concerns
- There is growing concern over
  - Cost of journals in era of constrained library budgets
  - Shift from paper to electronic access – you no longer get to keep your back issues

Call for "open access" to scientific research results

- Rationale: Most research publicly funded, yet reports of results copyrighted by publishers
  - If such information may be life-saving, it should be freely available
- Challenges: Production of information is not free and where do you draw the line with secondary publications
- Perspectives: Weiss, 2003; Healy, 2003
- Proposed legislation: Sabo bill will prohibit copyrighting of all US government-funded research (McLellan, 2003)

Open access publishing initiatives

- PubMed Central – pubmedcentral.gov
- BioMed Central (Hersh, 2001) – www.biomedcentral.com
- Public Library of Science (Butler, 2003) – www.plos.org
- Latter two bring publishing model full circle back to electronic equivalent of page charges in exchange for open access
  - Assumption that cost should be built into research budgets, with provisions for those unable to pay

Digital rights management (DRM)

- A huge area of interest, with many competing interests, proposals, and proprietary systems
- Martin et al. (2001) note that DRM in research and educational settings balance intellectual property protection with open and easy use

Conclusions

- IR systems have become "mainstream"
- Searching is an essential skill for knowledge workers and perhaps the rest of the world as well
- Basic searching is simple and easy to do
- Challenges remain in providing access to as much information as possible while preserving the incentive to produce it