

Grand Challenges for Information Retrieval in the Biomedical Domain

William Hersh
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
hersh@ohsu.edu
www.billhersh.info

References

- Anonymous (1990). *IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries*. Piscataway, NJ. IEEE Press.
- Anonymous (2001). Information for Health - A Strategy for Building the National Health Information Infrastructure. Washington, DC, National Committee on Vital and Health Statistics. <http://aspe.hhs.gov/sp/nhii/Documents/NHIIReport2001/default.htm>.
- Anonymous (2004). Washington D.C. Principles For Free Access to Science - A Statement from Not-for-Profit Publishers. Washington, DC, Washington D.C. Principles For Free Access to Science. <http://www.dcpinciples.org/statement.pdf>.
- Anonymous (2005). Physician Internet Use Statistics. <http://www.max.md/pdf/PhysicianInternetUseStatistics.pdf>.
- Anonymous (2006). Number of "Cyberchondriacs" - Adults Who Have Ever Gone Online for Health Information - Increases to an Estimated 136 Million Nationwide. Rochester, NY, Harris Interactive. http://www.harrisinteractive.com/harris_poll/index.asp?PID=686.
- 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.
- Butler, D. (2003). Who will pay for open access? *Nature*, 425: 554-555.
- Ely, J., Osheroff, J., et al. (1999). Analysis of questions asked by family doctors regarding patient care. *British Medical Journal*, 319: 358-361.
- Eysenbach, G. (2006). Citation advantage of open access articles. *PLoS Biology*, 4(5): e157.
- Fox, S. (2006). Online Health Search 2006. Washington, DC, Pew Internet & American Life Project. http://www.pewinternet.org/pdfs/PIP_Online_Health_2006.pdf.
- Gorman, P. (1995). Information needs of physicians. *Journal of the American Society for Information Science*, 46: 729-736.
- Haynes, R. (2001). Of studies, syntheses, synopses, and systems: the "4S" evolution of services for finding current best evidence. *ACP Journal Club*, 134: A11-A13.
- Hersh, W. (2001). The way of the future? Review of Biomed Central. *Nature*, 413: 680.
- Hersh, W. (2003). *Information Retrieval: A Health and Biomedical Perspective (Second Edition)*. New York. Springer-Verlag.
- Hersh, W. (2005). *Information Retrieval and Digital Libraries*, 237-275, in Chen, H., Fuller, S., Friedman, C. and Hersh, W., eds. *Medical Informatics: Knowledge Management and Data Mining in Biomedicine*. New York. Springer-Verlag.
- Hersh, W., Buckley, C., et al. (1994). OHSUMED: an interactive retrieval evaluation and new large test collection for research. *Proceedings of the 17th Annual International ACM*

- SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland. Springer-Verlag. 192-201.
- Hersh, W., Müller, H., et al. (2006). Advancing biomedical image retrieval: development and analysis of a test collection. *Journal of the American Medical Informatics Association*, 13: 488-496.
- 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.
- Lagoze, C. and VandeSompel, 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.
- Lawrence, S. (2001). Free online availability substantially increases a paper's impact. *Nature*, 411: 521.
- Salton, G. (1987). Historical note: the past thirty years in information retrieval. *Journal of the American Society for Information Science*, 38(5): 375-380.
- Schroter, S., Tite, L., et al. (2005). Perceptions of open access publishing: interviews with journal authors. *British Medical Journal*, 330: 756.
- Shortliffe, E. and Cimino, J., eds. (2006). *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*. New York, NY. Springer-Verlag.
- Srinivasan, P. (1996). Optimal document-indexing vocabulary for MEDLINE. *Information Processing and Management*, 32: 503-514.
- Swanson, D. (1988). Historical note: Information retrieval and the future of an illusion. *Journal of the American Society for Information Science*, 39: 92-98.
- Taylor, H. and Leitman, R. (2001). The Increasing Impact of eHealth on Physician Behavior. November 13, 2001.
- Wheeler, D., Barrett, T., et al. (2007). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 35: D5-D12.

Grand Challenges for Information Retrieval in the Biomedical Domain

William Hersh
 Professor and Chair
 Department of Medical Informatics & Clinical Epidemiology
 Oregon Health & Science University
 hersh@ohsu.edu
 www.billhersh.info



Overview

- Key attributes of information retrieval (IR) in biomedicine
- Grand challenges (among others)
 - Content – the right information for the right task
 - Indexing – metadata for Web content
 - Linkage – across multiple resources
 - Access – open access but protective of intellectual property

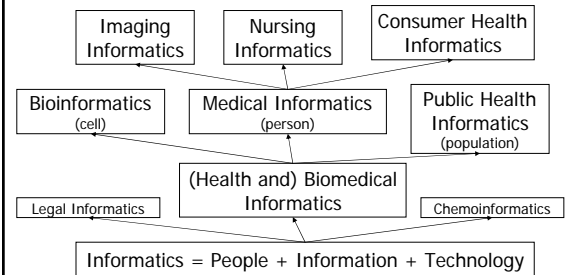
2



Information retrieval (IR) in biomedicine



How people from “informatics” see the world



(Adapted from Shortliffe, 2006)

4



The people from informatics who use IR systems and other apps

- From the “dimensions” in the National Health Infrastructure Report (NCVHS, 2001) report
 - Personal health – consumers, patients
 - Health care provider – physicians, nurses, others
 - Population health – public health officials
- Other groups
 - Researchers
 - Policy makers

5



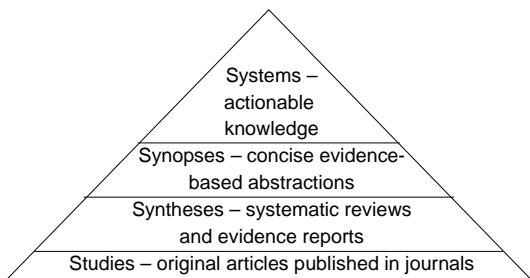
One view of clinical information (Hersh, 2003)

- 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

6



Another view from “evidence-based medicine”

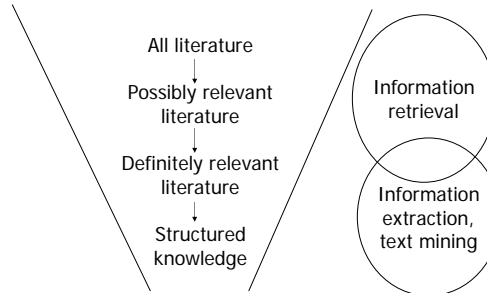


(adapted from Haynes, 2001)

7



IR also part of “knowledge discovery” (Hersh, 2005)



8



Search is becoming ubiquitous

- An estimated 80% of consumers who have used a search engine have searched for information related to personal health (Harris Interactive, 2006; Fox, 2006)
 - In other words, about 136 million Americans seek health information on Internet
- Most clinicians are connected
 - About 98% of US physicians use the Internet and half use PDAs (Physician Internet Use Statistics, 2005)
 - Is used more by those who are busier seeing patients (Taylor and Leitman, 2001)

9



But new problems have emerged



10



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 learning how to use search systems better, but are not major research challenges
 - There is a major role for librarians, informaticians, and others

11



Grand challenges for biomedical IR

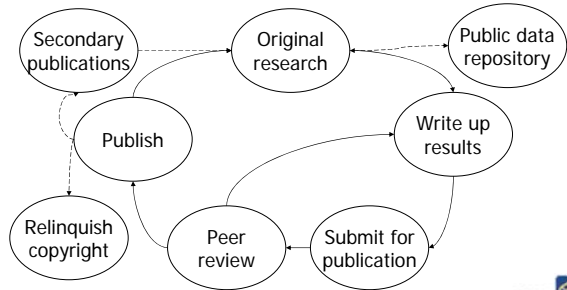
- Covered in this talk
 - Content – the right information for the right task
 - Indexing – metadata for Web content
 - Linkage – across multiple resources
 - Access – open access but protective of intellectual property
- Others of great interest, for another day
 - Evaluation – best measures, meaningful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

12



Content

The life-cycle of knowledge-based information



14

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

15

Most important issues concerning content

- Need 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

16

How do we produce refined knowledge?

- Understanding users' questions, e.g., Gorman, 1995; Ely, 1999; and others
- Mass production is challenging
 - Academic rewards for tasks other than knowledge production 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

17

Indexing

The age-old debate on human vs. automated indexing

- Manifested in earlier times by Gerald Salton vs. Don Swanson dating back to 1980s
- Salton (1987)
 - Indexers too inconsistent
 - Word-indexed systems consistently as effective as human-indexed ones
- Swanson (1988)
 - Machines cannot duplicate human judgement
 - Humans good at recognizing focus of documents

19



But manual indexing in biomedicine does add value

- In a variety of places, value of Medical Subject Headings (MeSH) indexing has shown value
 - Cost of indexing MEDLINE is large (\$30-40M) but not substantial relative to other expenditures of US government
- Hersh, 1994 – OHSUMED collection showed 10% performance improvement with use of MeSH indexing for MEDLINE
- Srinivasan, 1996 – showed additional value for MeSH terms in query expansion
- Various approaches and tasks in TREC Genomics Track have shown benefit for synonym expansion and other aspects of MeSH

20



Other forms of metadata usage also appear to provide value

- Annotation of genes and their functions, e.g., Gene Ontology (GO)
- We still do not know how to annotate non-textual objects, such as images (e.g., ImageCLEF medical task, Hersh, 2006)
- If (a big if!) semantic Web ever becomes reality, some and perhaps a great deal of manual or semi-automated indexing will be required

21



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

23



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

24



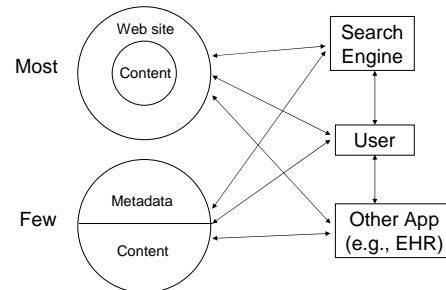
Impediments for others

- 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

25



The current problem: most information is in *silos*



26



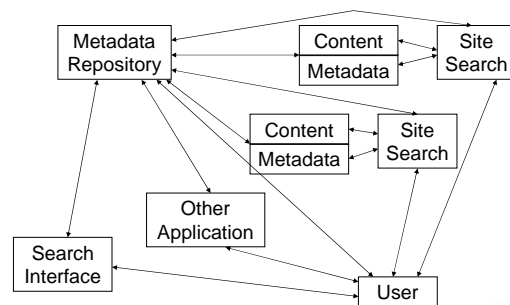
Overcoming the impediments: Interoperability

- 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

27



From silos to interoperability



28



How might we achieve this?

- Possible 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 is OAI Repository Explorer: <http://re.cs.ucl.ac.za/>

29



Are there any good examples of integrated resources?

- Yes, from the genomics community
- Databases of National Center for Biotechnology Information (NCBI) are linked in Entrez (Wheeler, 2007; <http://www.ncbi.nlm.nih.gov/Entrez/>) and include
 - Literature
 - Nucleotide and protein sequences
 - Protein structures
 - Textbooks and other textual resources
 - Genomes and map
- Which leads us to issues of access...

30



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 (in the case of biomedical research), should it be freely available?
 - Freely available articles more likely to be cited (Lawrence, 2001; Eysenbach, 2006)
- Challenges: Production of information is not free
- Perspectives: Authors are sympathetic but have higher concerns than free access, i.e., publication in prestigious journals (Schroter, 2006)

Open access publishing initiatives in biomedicine

- PubMed Central – pubmedcentral.gov
- BioMed Central (Hersh, 2001) – www.biomedcentral.com
- Public Library of Science (Butler, 2003) – www.plos.org
- Pushback from non-profit publishers – *Washington D.C. Principles For Free Access to Science* (2004)

Conclusions

- IR systems, especially in biomedicine, 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 creating and providing access to the right information for the right task while preserving the incentive to produce it