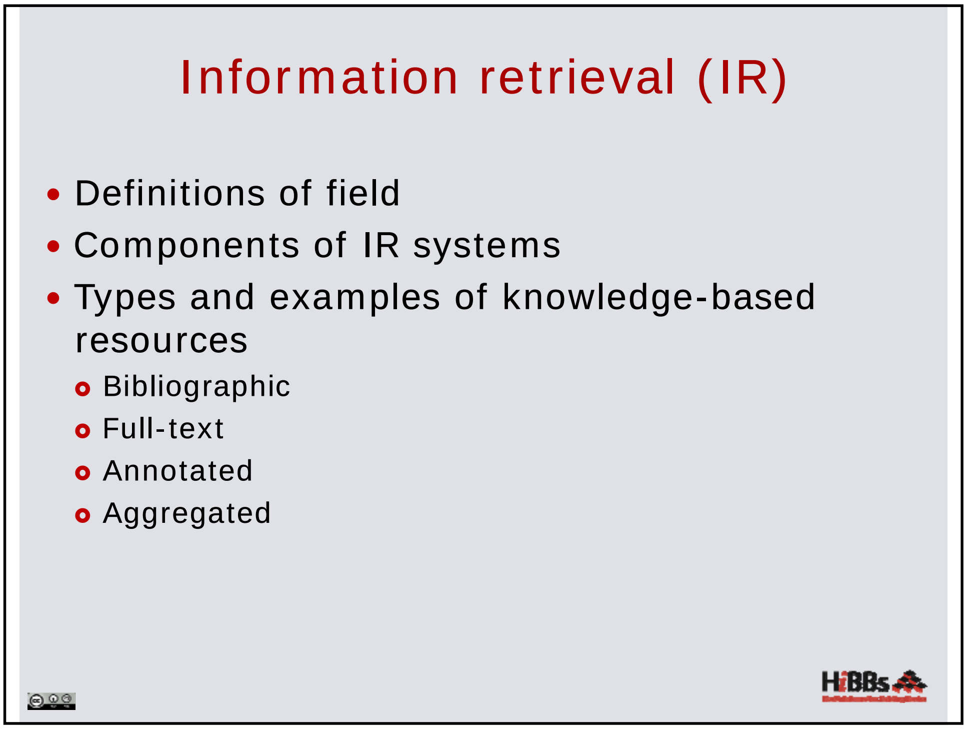
Information Retrieval: Access to Knowledge-Based Resources

WILLIAM HERSH, MD
OREGON HEALTH & SCIENCE UNIVERSITY



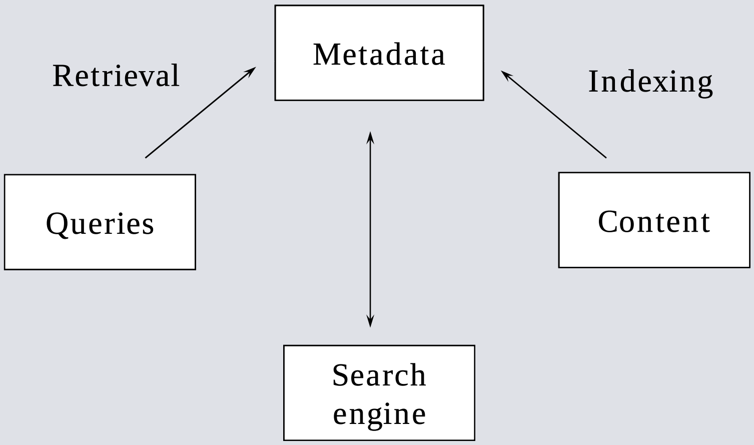
Information retrieval (IR)

- Definitions of field
- Components of IR systems
- Types and examples of knowledge-based resources
  - Bibliographic
  - Full-text
  - Annotated
  - Aggregated

Information retrieval (IR)

- Field concerned with organization and retrieval of knowledge-based information
  - Focuses mainly on textual information, but multimedia (e.g., images, sounds, video, etc.) and more complex databases are increasingly a part
  - Historically not focused on patient-based information, but this is changing too
- IR is also sometimes called “search”
  - Is probably most prevalent activity on Web, by clinicians and patients alike

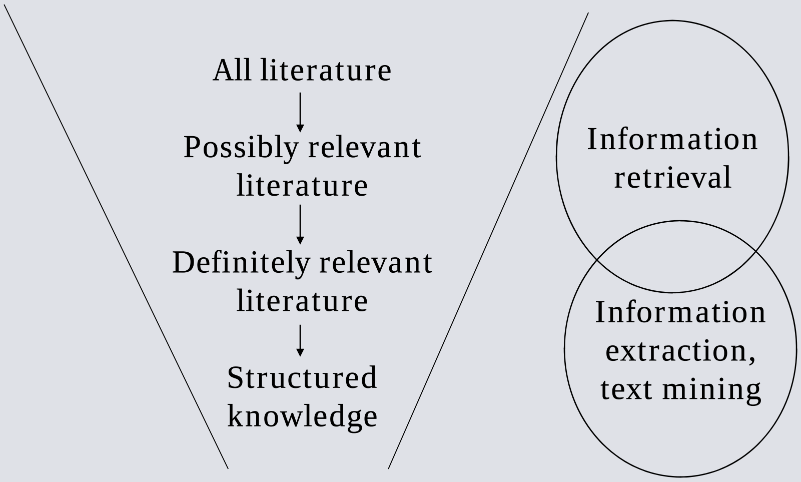
Components of IR systems

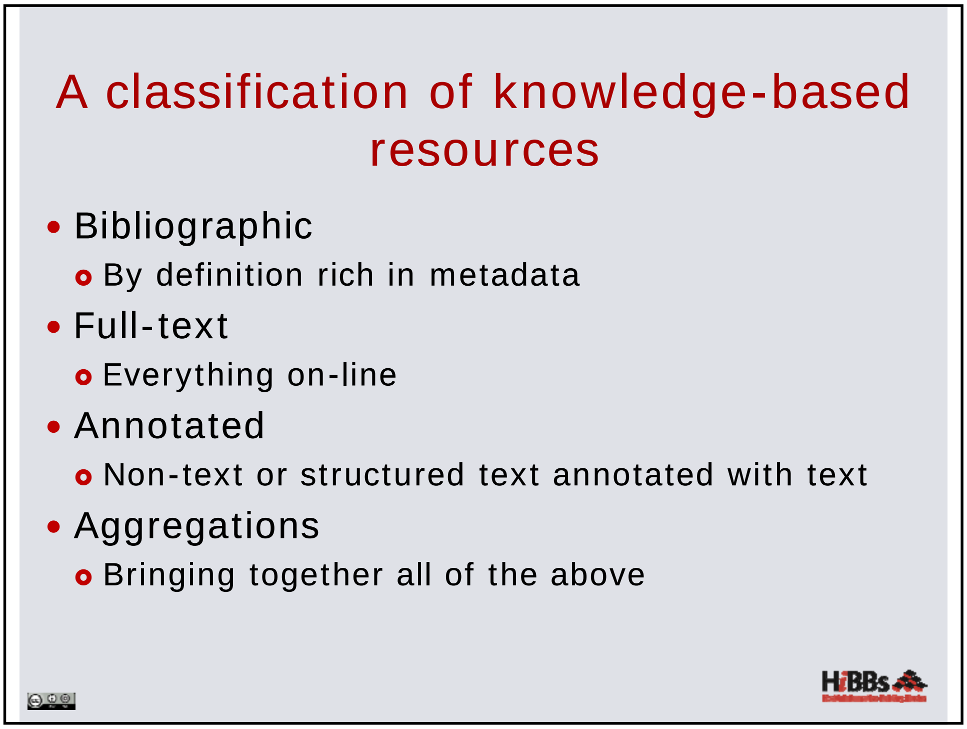


The intellectual tasks of IR

- Indexing
  - Assigning metadata to content items
  - Can assign
    - Subjects (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

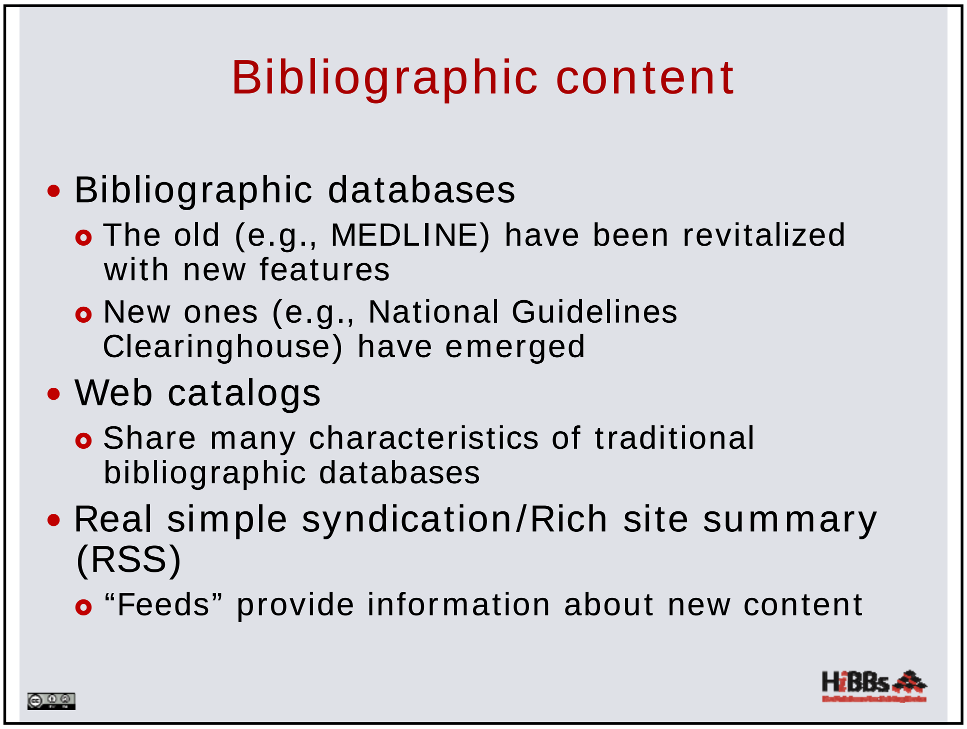
IR also a growing part of "knowledge discovery"


All literature
Possibly relevant literature
Definitely relevant literature
Structured knowledge
Information retrieval
Information extraction, text mining



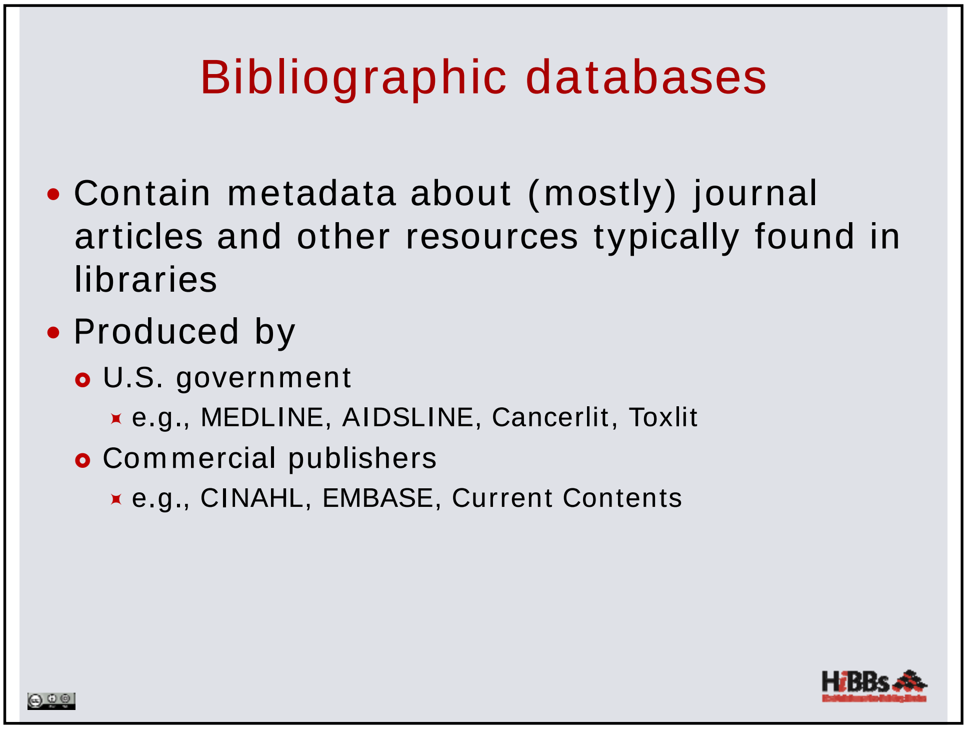
A classification of knowledge-based resources

- Bibliographic
  - By definition rich in metadata
- Full-text
  - Everything on-line
- Annotated
  - Non-text or structured text annotated with text
- Aggregations
  - Bringing together all of the above



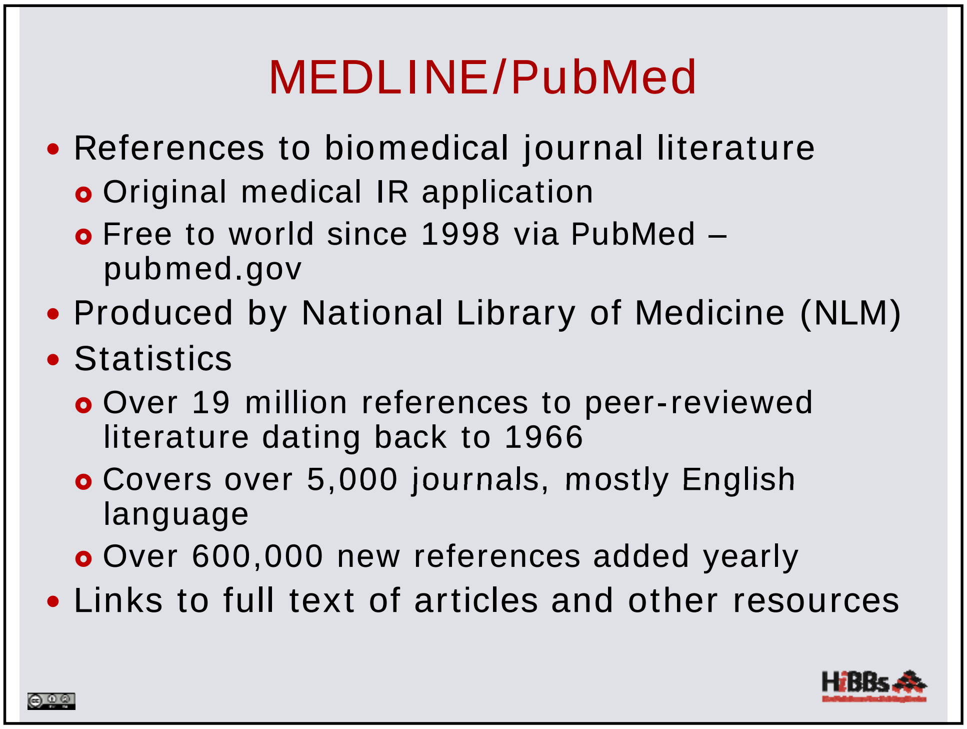
Bibliographic content

- Bibliographic databases
  - The old (e.g., MEDLINE) have been revitalized with new features
  - New ones (e.g., National Guidelines Clearinghouse) have emerged
- Web catalogs
  - Share many characteristics of traditional bibliographic databases
- Real simple syndication/Rich site summary (RSS)
  - “Feeds” provide information about new content



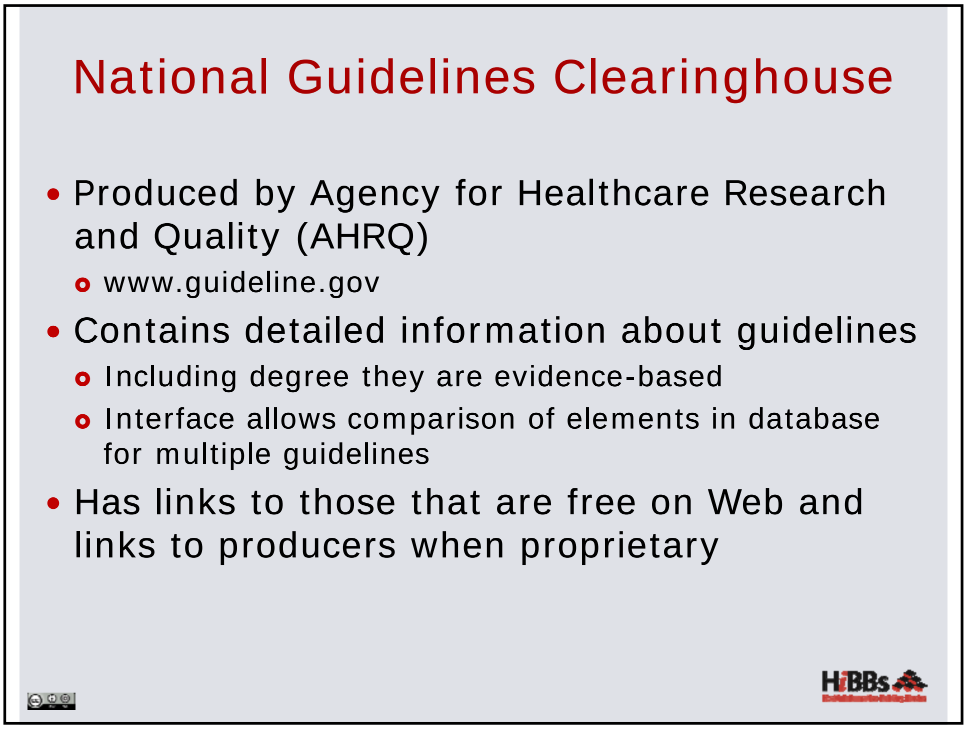
Bibliographic databases

- Contain metadata about (mostly) journal articles and other resources typically found in libraries
- Produced by
  - U.S. government
    - e.g., MEDLINE, AIDSLINE, Cancerlit, Toxlit
  - Commercial publishers
    - e.g., CINAHL, EMBASE, Current Contents



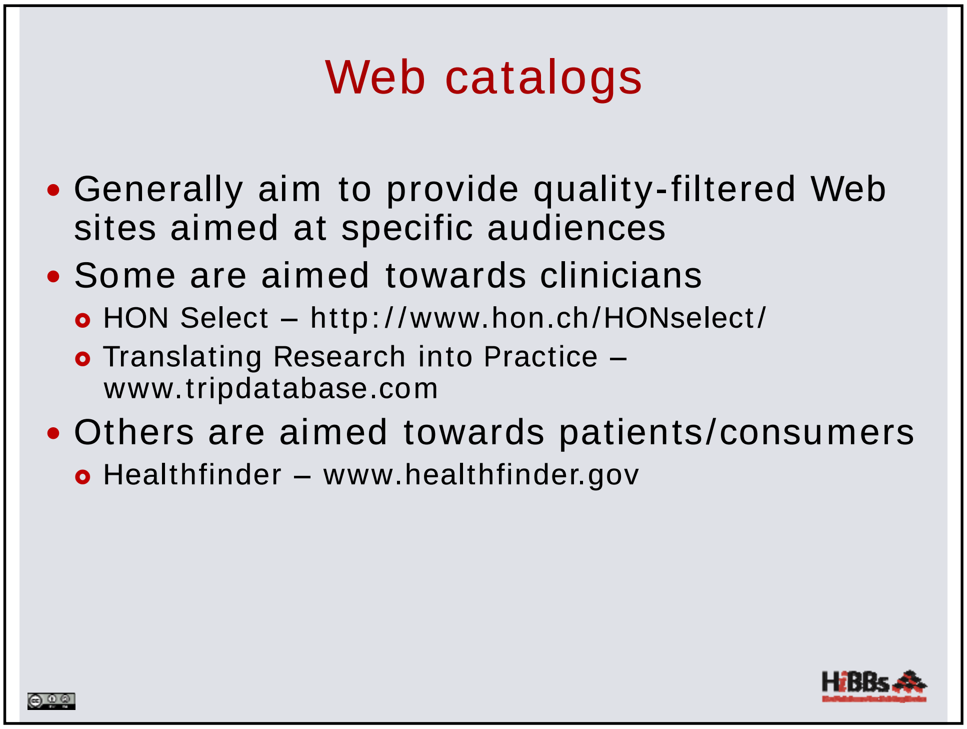
MEDLINE/PubMed

- References to biomedical journal literature
  - Original medical IR application
  - Free to world since 1998 via PubMed – pubmed.gov
- Produced by National Library of Medicine (NLM)
- Statistics
  - Over 19 million references to peer-reviewed literature dating back to 1966
  - Covers over 5,000 journals, mostly English language
  - Over 600,000 new references added yearly
- Links to full text of articles and other resources



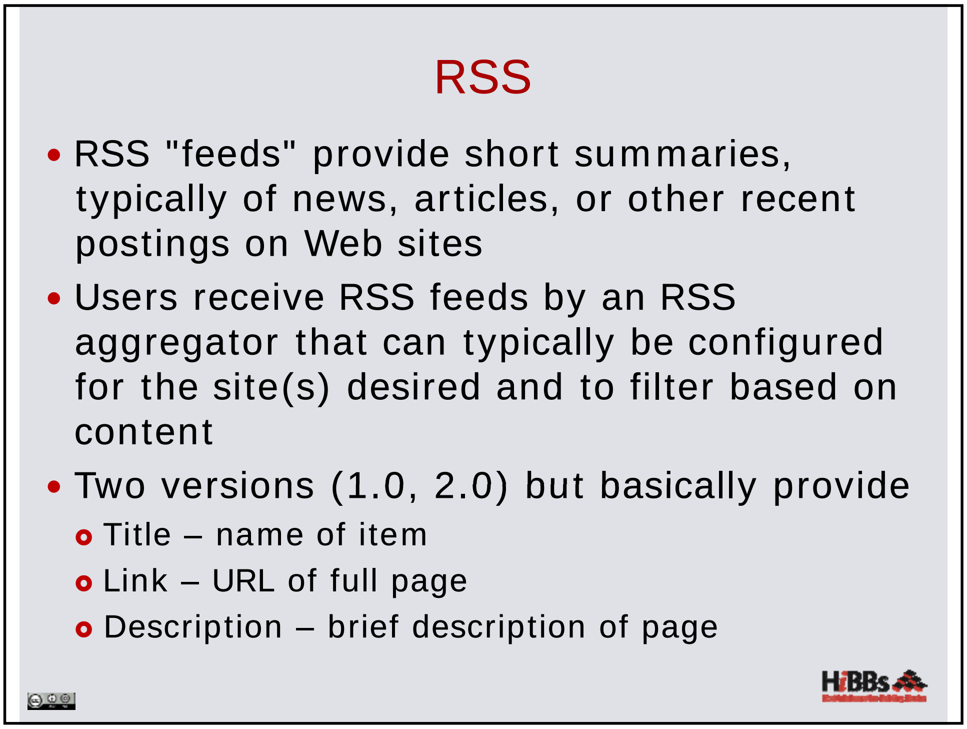
National Guidelines Clearinghouse

- Produced by Agency for Healthcare Research and Quality (AHRQ)
  - www.guideline.gov
- Contains detailed information about guidelines
  - Including degree they are evidence-based
  - Interface allows comparison of elements in database for multiple guidelines
- Has links to those that are free on Web and links to producers when proprietary



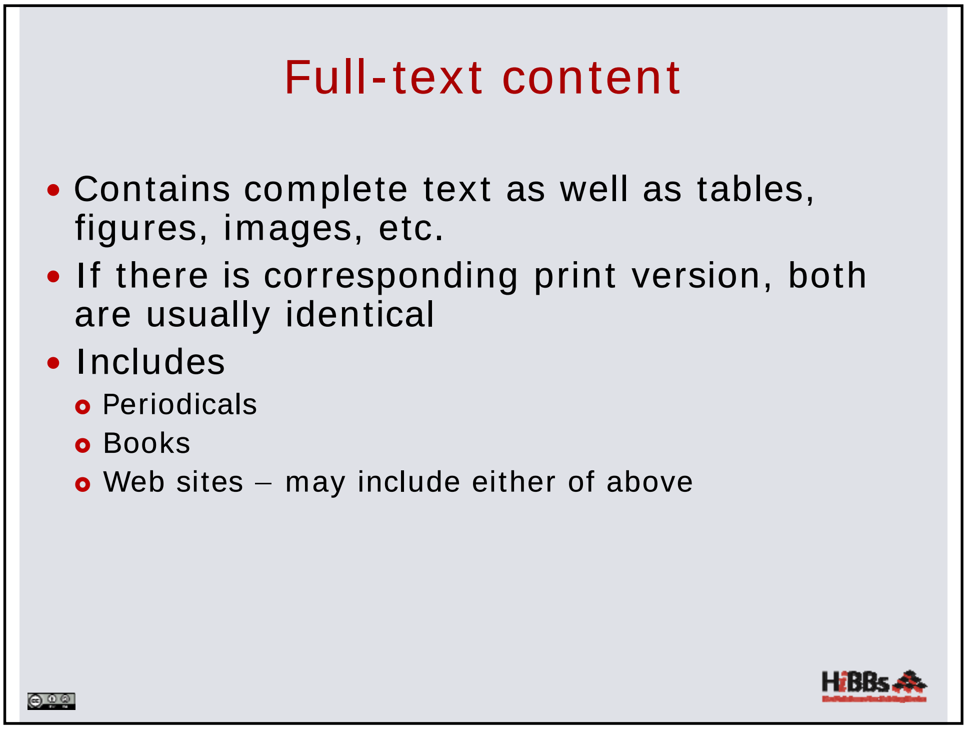
Web catalogs

- Generally aim to provide quality-filtered Web sites aimed at specific audiences
- Some are aimed towards clinicians
  - HON Select – http://www.hon.ch/HONselect/
  - Translating Research into Practice – www.tripdatabase.com
- Others are aimed towards patients/consumers
  - Healthfinder – www.healthfinder.gov



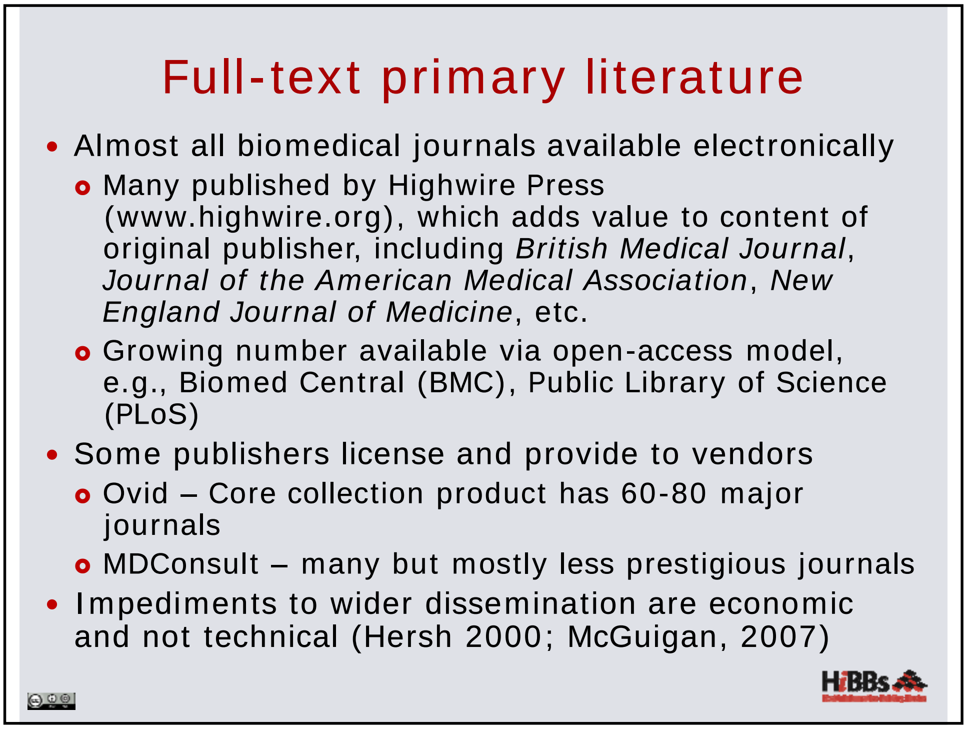
RSS

- RSS "feeds" provide short summaries, typically of news, articles, or other recent postings on Web sites
- Users receive RSS feeds by an RSS aggregator that can typically be configured for the site(s) desired and to filter based on content
- Two versions (1.0, 2.0) but basically provide
  - Title – name of item
  - Link – URL of full page
  - Description – brief description of page



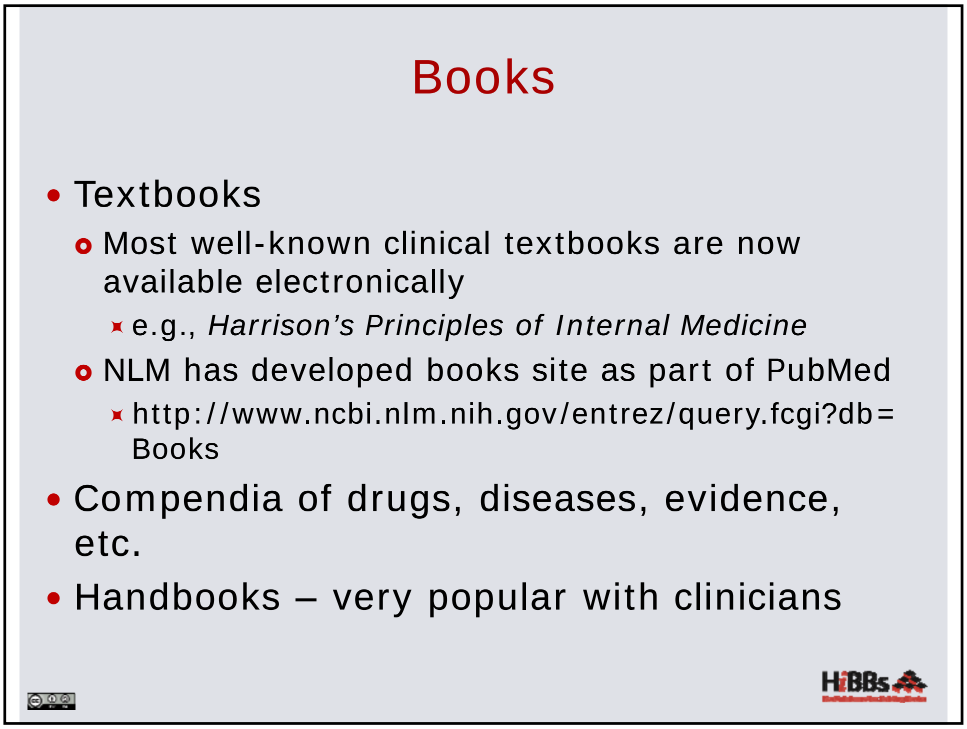
Full-text content

- Contains complete text as well as tables, figures, images, etc.
- If there is corresponding print version, both are usually identical
- Includes
  - Periodicals
  - Books
  - Web sites – may include either of above



Full-text primary literature

- Almost all biomedical journals available electronically
  - Many published by Highwire Press (www.highwire.org), which adds value to content of original publisher, including *British Medical Journal*, *Journal of the American Medical Association*, *New England Journal of Medicine*, etc.
  - Growing number available via open-access model, e.g., Biomed Central (BMC), Public Library of Science (PLoS)
- Some publishers license and provide to vendors
  - Ovid – Core collection product has 60-80 major journals
  - MDConsult – many but mostly less prestigious journals
- Impediments to wider dissemination are economic and not technical (Hersh 2000; McGuigan, 2007)

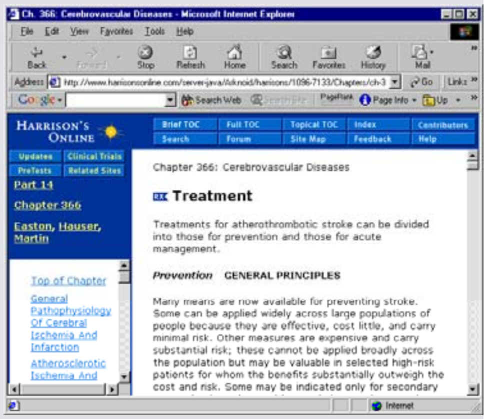


Books

- Textbooks
  - Most well-known clinical textbooks are now available electronically
    - e.g., *Harrison's Principles of Internal Medicine*
  - NLM has developed books site as part of PubMed
    - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books
- Compendia of drugs, diseases, evidence, etc.
- Handbooks – very popular with clinicians

Value added for electronic books

- Multimedia, e.g., skin lesions, shuffling gait of Parkinson's Disease, etc.
- Bundling of multiple books
- Can be updated in between "editions"
- Linkage to other information, e.g., to references, self-assessments, updates, other resources, etc.


Ch. 366: Cerebrovascular Diseases - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail
Address http://www.harrisonsonline.com/server-java/Arknoid/harrisons/1096-7133/Chapters/ch-3
Go Links
Search Web PageRank Page Info Up
HARRISON'S ONLINE
Brief TOC Full TOC Topical TOC Index Contributors
Search Forum Site Map Feedback Help
Updates Clinical Trials
PreTests Related Sites
Part 14
Chapter 366
Easton, Hauser, Martin
Top of Chapter
General Pathophysiology Of Cerebral Ischemia And Infarction
Atherosclerotic Ischemia And
Internet

Chapter 366: Cerebrovascular Diseases

Treatment

Treatments for atherothrombotic stroke can be divided into those for prevention and those for acute management.

Prevention GENERAL PRINCIPLES

Many means are now available for preventing stroke. Some can be applied widely across large populations of people because they are effective, cost little, and carry minimal risk. Other measures are expensive and carry substantial risk; these cannot be applied broadly across the population but may be valuable in selected high-risk patients for whom the benefits substantially outweigh the cost and risk. Some may be indicated only for secondary

Web sites

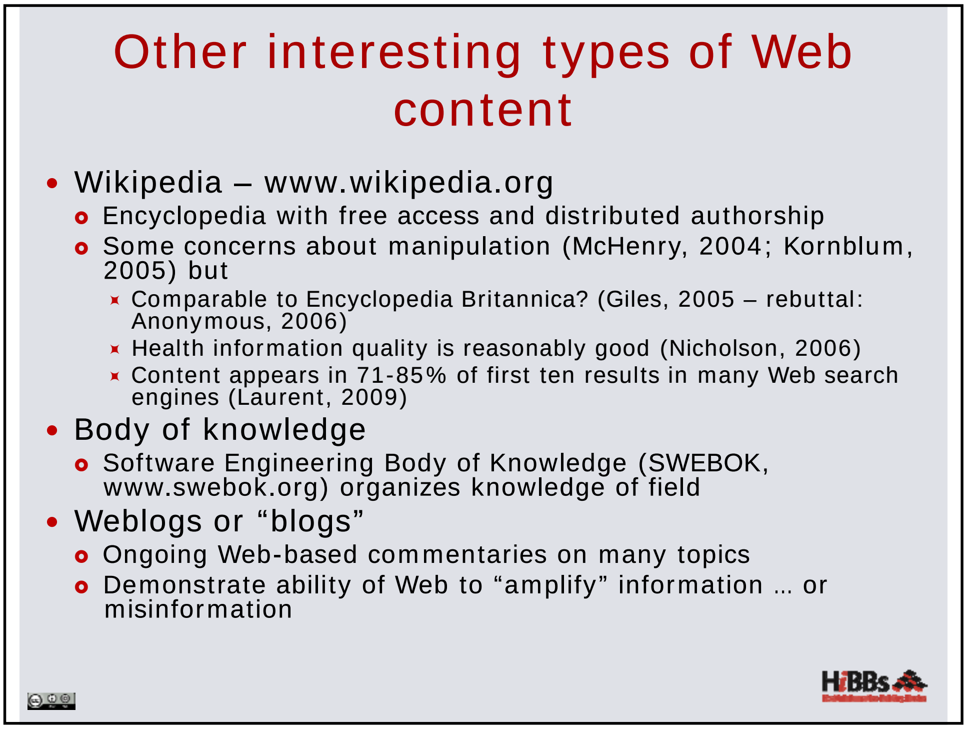
- Defined more narrowly here to refer to coherent collections of information on Web
- Usually take advantage of Web features, such as linking, multimedia

Some notable full-text content on Web sites

- Government agencies
  - CancerNet – from National Cancer Institute
    - www.cancer.gov
  - Centers for Disease Control – travel and infection information
    - www.cdc.gov
    - http://www.cdc.gov/travel/
  - Other NIH institutes, e.g., National Heart, Lung, and Blood Institute (NHLBI)
    - www.nhlbi.nih.gov

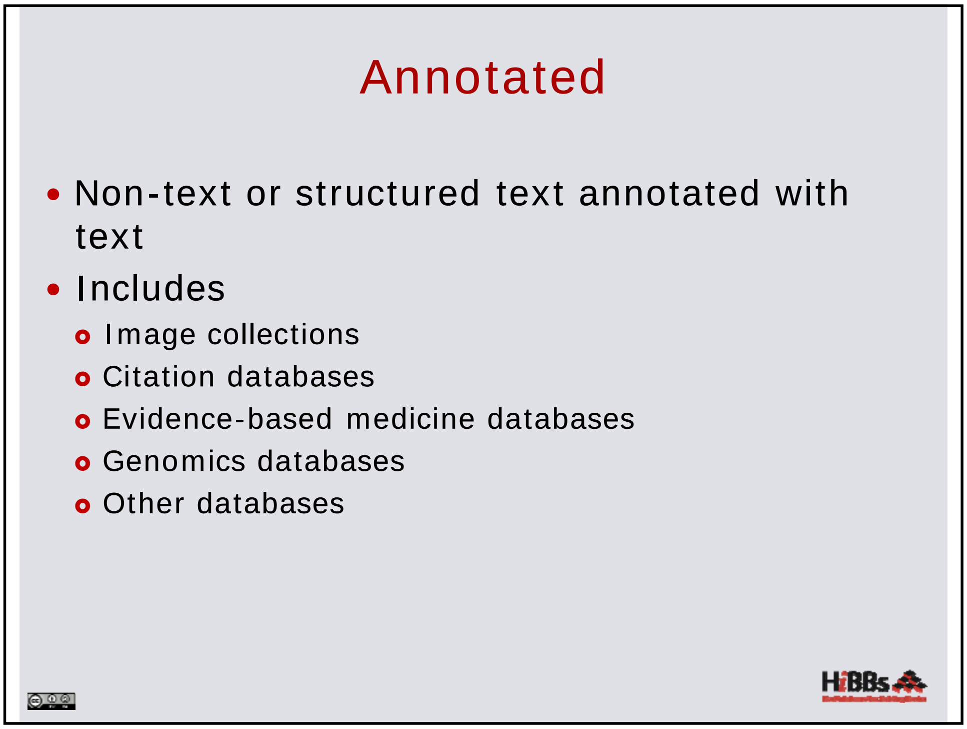
Full-text Web sites (cont.)

- Physician-oriented medical news and overviews, e.g.,
  - Medscape – www.medscape.com
  - PEPID – www.pepid.com
  - Many professional societies provide to members
- Patient/consumer-oriented, e.g.,
  - Intelihealth – www.intelihealth.com
  - NetWellness – www.netwellness.com



Other interesting types of Web content

- Wikipedia – www.wikipedia.org
  - Encyclopedia with free access and distributed authorship
  - Some concerns about manipulation (McHenry, 2004; Kornblum, 2005) but
    - Comparable to Encyclopedia Britannica? (Giles, 2005 – rebuttal: Anonymous, 2006)
    - Health information quality is reasonably good (Nicholson, 2006)
    - Content appears in 71-85% of first ten results in many Web search engines (Laurent, 2009)
- Body of knowledge
  - Software Engineering Body of Knowledge (SWEBOK, www.swebok.org) organizes knowledge of field
- Weblogs or “blogs”
  - Ongoing Web-based commentaries on many topics
  - Demonstrate ability of Web to “amplify” information … or misinformation



Annotated

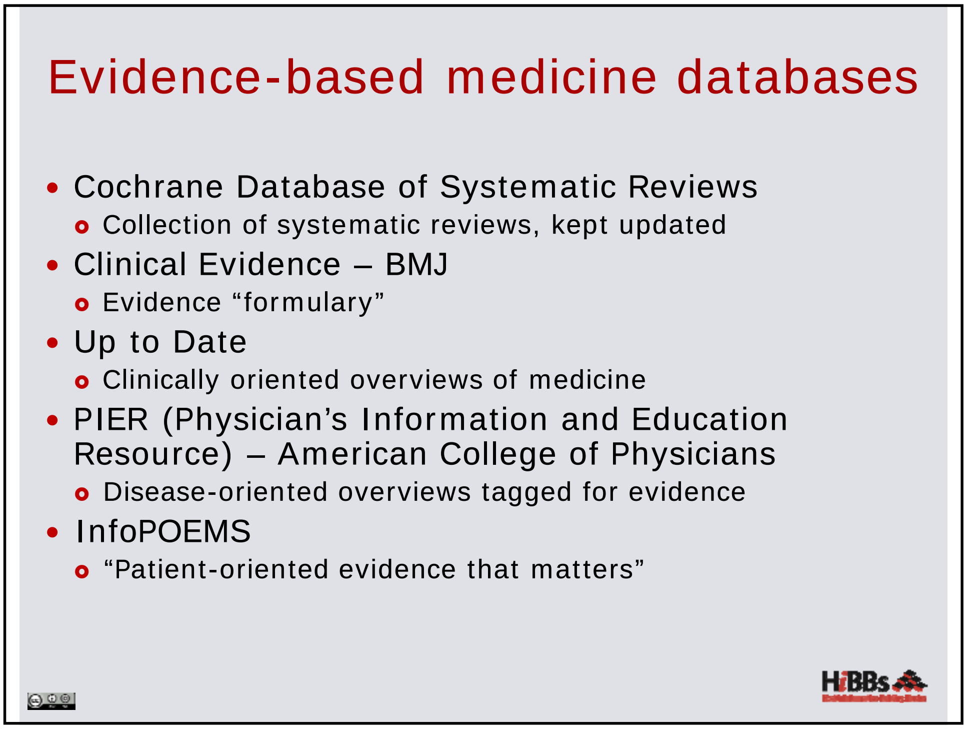
- Non-text or structured text annotated with text
- Includes
  - Image collections
  - Citation databases
  - Evidence-based medicine databases
  - Genomics databases
  - Other databases

Image collections

- Most prominent in the "visual" medical specialties, such as radiology, pathology, and dermatology
- Well-known collections include
  - Visible Human – http://www.nlm.nih.gov/research/visible/visible_human.html
  - BrighamRad – http://harvardscience.harvard.edu/directory/programs/brighamrad
  - WebPath – http://library.med.utah.edu/WebPath/webpath.html
  - More pathology – PEIR, www.peir.net
  - DermIS – www.dermis.net
- Many have associated text, which assists with indexing and retrieval

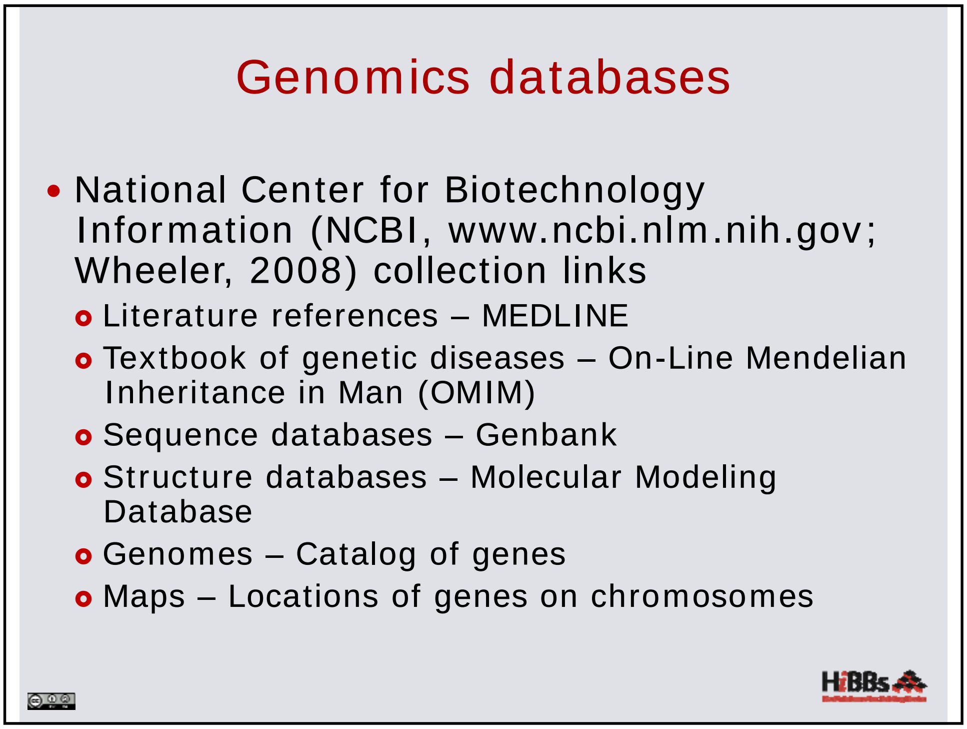
Citation databases

- *Science Citation Index* and *Social Science Citation Index*
  - Database of journal articles that have been cited by other journal articles
  - Now part of a package called *Web of Science*, which itself is part of larger project, *Web of Knowledge* (Thomson-Reuters)
    - isiwebofknowledge.com
- SCOPUS – info.scopus.com
- Google Scholar – scholar.google.com



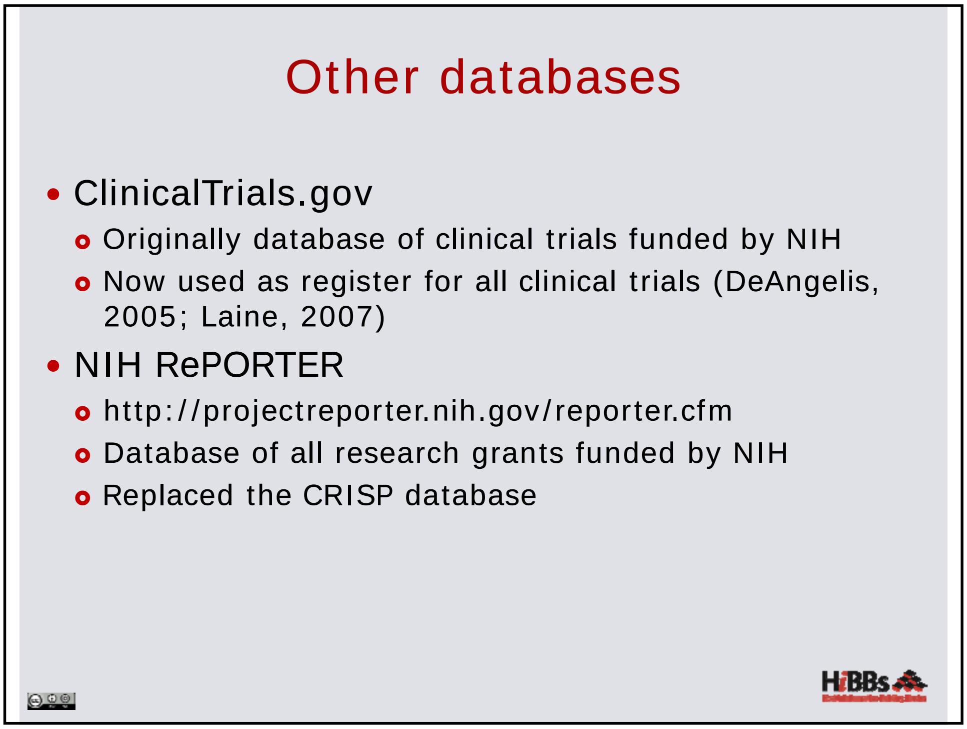
Evidence-based medicine databases

- Cochrane Database of Systematic Reviews
  - Collection of systematic reviews, kept updated
- Clinical Evidence – BMJ
  - Evidence “formulary”
- Up to Date
  - Clinically oriented overviews of medicine
- PIER (Physician’s Information and Education Resource) – American College of Physicians
  - Disease-oriented overviews tagged for evidence
- InfoPOEMS
  - “Patient-oriented evidence that matters”



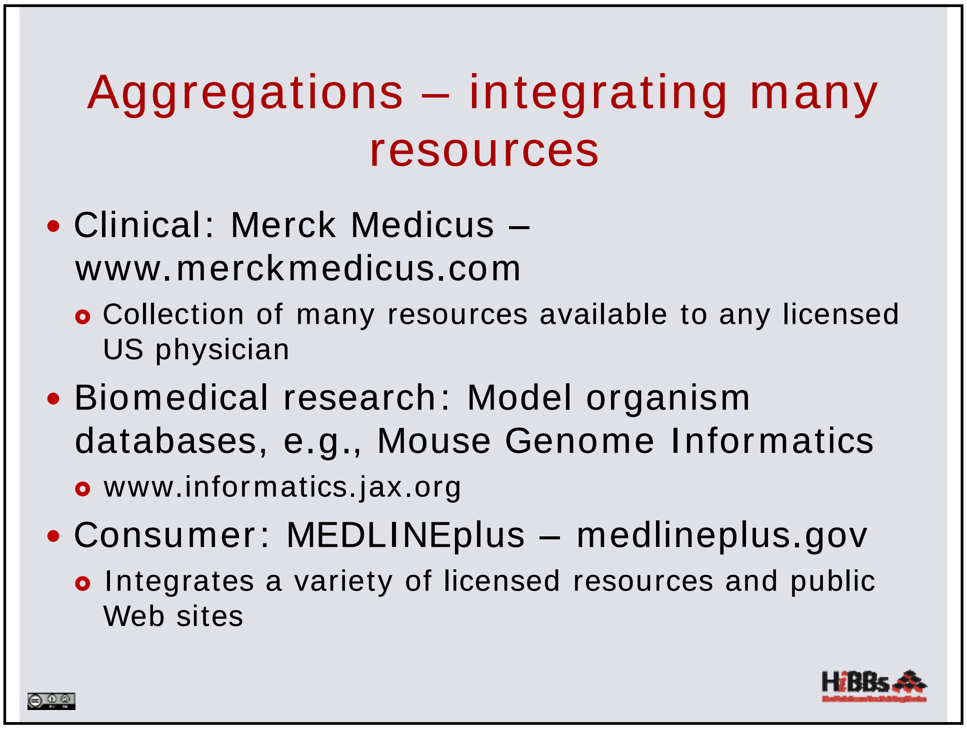
Genomics databases

- National Center for Biotechnology Information (NCBI, www.ncbi.nlm.nih.gov; Wheeler, 2008) collection links
  - Literature references – MEDLINE
  - Textbook of genetic diseases – On-Line Mendelian Inheritance in Man (OMIM)
  - Sequence databases – Genbank
  - Structure databases – Molecular Modeling Database
  - Genomes – Catalog of genes
  - Maps – Locations of genes on chromosomes



Other databases

- ClinicalTrials.gov
  - Originally database of clinical trials funded by NIH
  - Now used as register for all clinical trials (DeAngelis, 2005; Laine, 2007)
- NIH RePORTER
  - http://projectreporter.nih.gov/reporter.cfm
  - Database of all research grants funded by NIH
  - Replaced the CRISP database



Aggregations – integrating many resources

- Clinical: Merck Medicus – www.merckmedicus.com
  - Collection of many resources available to any licensed US physician
- Biomedical research: Model organism databases, e.g., Mouse Genome Informatics
  - www.informatics.jax.org
- Consumer: MEDLINEplus – medlineplus.gov
  - Integrates a variety of licensed resources and public Web sites