4.2 Information Retrieval Content

What is Biomedical and Health Informatics? - http://informatics.health/
William Hersh, MD, FACMI, FAMIA, FIAHSI

Classification of knowledge-based content

- Bibliographic
  - By definition rich in metadata
- Full-text
  - Everything on-line
- Semi-structured
  - Non-text or structured text annotated with text
- Aggregations
  - Bringing together all of the above
- These categories are somewhat fuzzy, and increasing numbers of resources have more than one type

Bibliographic content

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

Bibliographic databases

- Contain metadata about (mostly) journal articles and other resources typically found in libraries
- Produced by
  - U.S. government – most produced by National Library of Medicine (NLM)
    - e.g., MEDLINE, omics information, etc.
  - Commercial publishers, e.g.,
    - EMBASE – part of larger SciVal
    - CINAHL – Cumulative Index to Nursing and Allied Health Literature
    - ACM Guide to Computing Literature – computer science and related areas
    - Google Scholar

MEDLINE

- References to biomedical journal literature
- Original medical IR database – system for searching MEDLINE launched in 1971 with literature maintained in MEDLARS system dating back to 1966
  - Name derives from MEDLARS On-Line – MEDLINE
- NLM maintains policies for addition and maintenance
- Free to world since 1997 via PubMed
  - Now with links to full text of articles and other resources
  - Beginning to add preprints, e.g., BioRxiv and MedRxiv
  - Contains some additional content not in MEDLINE, e.g., PubMed Central and books

MEDLINE statistics

	FY2023	FY2022	FY2021	FY2020	FY2019	FY2018
MEDLINE Citations Indexed (Annual)	1,279,327	1,369,611	1,291,807	952,919	956,390	904,636
MEDLINE Citations Cumulative Total	30,966,708	29,807,639	28,444,654	27,149,277	26,196,358	25,239,968
MEDLINE Journal Titles	5,294	5,282	5,282	5,274	5,243	5,251
PubMed Citations (Annual)	1,567,478	1,714,780	1,733,089	1,514,199	1,366,447	1,329,148
PubMed Citations Cumulative Total	36,555,430	34,693,538	33,136,289	31,563,992	30,178,674	28,934,389
PubMed Searches	3.66 Billion	2.58 Billion	2.57 Billion	3.3 Billion	3.1 Billion	3.3 Billion
Web/Interactive	1.498 Billion	1.283 Billion	1.186 Billion	1.076 Billion	896 Million	831 Million
Script/E-Utilities	2.166 Billion	1.303 Billion	1.391 Billion	2.2 Billion	2.2 Billion	2.5 Billion

OHSU

ECRI Guidelines Trust

- Contains detailed information about guidelines
  - Including degree they are evidence-based
  - Interface allows comparison of elements in database for multiple guidelines
- Links to those free on Web and to producers when proprietary
- Successor to Agency for Healthcare Research and Quality (AHRQ) National Guidelines Clearinghouse

Web catalogs

- Generally aim to provide quality-filtered Web sites aimed at specific audiences
  - Distinction between catalogs and sites blurry
- Some are aimed towards clinicians
  - Translating Research into Practice (TRIP)
- Others are aimed towards patients/consumers
  - MedlinePlus – part of larger consumer health site

RSS – bibliographic feeds (Target, 2019)

- RSS “feeds” provide short summaries, typically of news, journal 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
  - Work as standalone, in Web browsers, in email clients, etc.
- Forked into different versions but basically provided
  - 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

OHSU

Full-text primary literature

- Almost all biomedical journals available electronically
- Many initially published by Highwire Press, which added value to content of original publisher
- Now also published by leading commercial scientific publishers, e.g., Elsevier, Kluwer, Springer, etc.
- Growing number available via open-access model, e.g., Biomed Central (BMC), Public Library of Science (PLoS)

Full-text literature before publication

- Repository of full-text papers funded by NIH research
  – PubMed Central (PMC)
- Preprint servers – some journals maintain but also general sites
  – arXiv
  – Biology – bioRxiv
  – Medicine – medRxiv

Books

- Textbooks
  - Most well-known clinical textbooks are now available electronically
    - e.g., Harrison's Principles of Internal Medicine
  - NLM Bookshelf
- Compendia of drugs, diseases, evidence, etc.
- Handbooks – very popular with clinicians
- Many of above are bundled into aggregations by publishers
  - e.g., Access Medicine (McGraw-Hill), Elsevier, Kluwer
  - Also increasingly published on mobile devices

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.

OHSU

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
- Increasingly integrated with other resources and available on different platforms (e.g., integrated into electronic health records [EHRs], on smartphones, etc.)

Some notable full-text content on Web sites

- US government agencies
  - National Cancer Institute – Cancer.gov
  - Centers for Disease Control – travel and infection information
    - Health Topics
    - Travel
  - Other NIH institutes, e.g., National Heart, Lung, and Blood Institute (NHLBI)

Full-text Web sites (cont.)

- Physician-oriented medical news and overviews, e.g.,
  - Medscape
  - Many professional societies provide to members, e.g., American College of Physicians (ACP)
  - Some devoted to specialties, e.g., Orthopaedia
- Patient/consumer-oriented, e.g.,
  - MayoClinic.org
  - WebMD
- Many mobile apps provide health information, e.g.,
  - Epocrates
  - WebMD app for consumers

Other types of Web content

- Wikipedia
  - Encyclopedia with free access and distributed authorship
  - Medical content often retrieved in general Web searches (Laurent, 2009)
  - Making attempt to improve quality of medical content (Heilman, 2013; Shafee, 2017; Azzam, 2017)
- Body of knowledge
  - Software Engineering Body of Knowledge (SWEBOK) organizes knowledge of field
- Social media and beyond – X/Twitter, Facebook, LinkedIn, BlueSky, etc.

Semi-structured content

- Non-text or structured text annotated with text
- Includes
  - Image collections
  - Citation databases
  - Evidence-based medicine databases
  - Clinical decision support
  - Computable biomedical knowledge
  - Omics databases
  - Data repositories
  - Other databases

Image collections

- Most prominent in the “visual” medical specialties, such as radiology, pathology, and dermatology
- Come and go, but well-known collections include
  - Open-I
  - Radiopedia
  - Visible Human
  - WebPath
  - More pathology – PEIR
  - DermIS
  - More dermatology – VisualDx, also a decision-support system
- 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
- SCOPUS
- Google Scholar

Evidence-based medicine content

- Cochrane Database of Systematic Reviews
  - Collection of systematic reviews, kept updated
- Evidence “formularies”
  - JAMAevidence
- AHRQ Evidence Reports
- Many resources part of aggregations

OHSU

Clinical decision support (CDS)

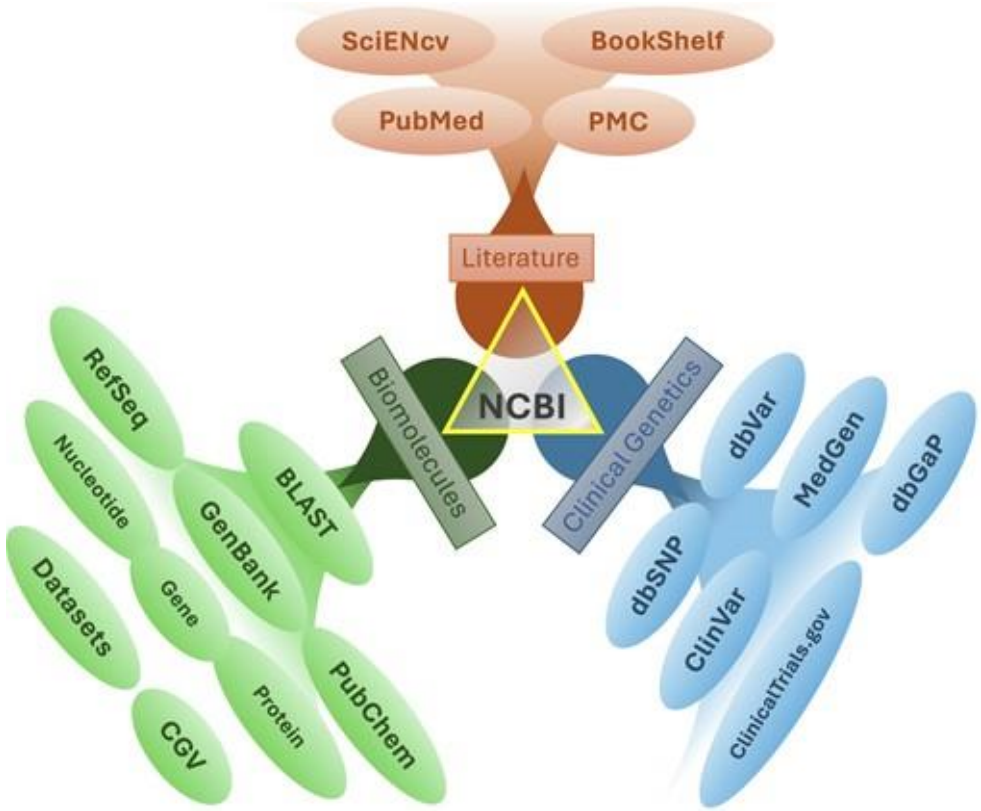
- Content used in CDS systems, usually part of EHRs
  - Order sets (usually “evidence-based”)
  - CDS rules
  - Health/disease management templates
- Growing and evolving commercial market for such tools, especially as EHR adoption increases; leaders include
  - Zynx
  - Provation
  - EHR vendors themselves and partners

Computational biomedical knowledge (CBK)

- Digital objects representing biomedical knowledge in machine-interpretable structures
- Requires technical infrastructure to support and maintain (McCusker, 2023)
- Metadata categories describe type, domain, location, technical characteristics, authorization, and evidence basis (Alper, 2022)
- Small but growing number of CBK repositories (Platt, 2023)

Omics databases

- National Center for Biotechnology Information (NCBI; Sayers, 2025) collection links
  - Literature references – MEDLINE
  - Textbook of genetic diseases – On-Line Mendelian Inheritance in Man (OMIM)
  - Sequence databases – Genbank
  - Structure databases
  - Genomes – catalogs of genes
  - Clinical associations – ClinVar
- More in bioinformatics unit…



Data repositories

- With growing publication of data, need methods to discover and access metadata
- Journal articles and reports with published data may
  - Make data available as supplement on journal site
  - Provide instructions to request from author
- One of largest publishers of data is US government, e.g.,
  - Data.gov
  - NLM Dataset Catalog
  - NCBI Datasets (O'Leary, 2024)
- Many other sites for data; some include (or designed for) code
  - Github – versioning site for code; also used for data sets
  - Hugging Face – datasets and models for machine learning

Other databases

- ClinicalTrials.gov
  - Originally database of clinical trials funded by NIH
  - Also used as register for clinical trials (DeAngelis, 2005; Laine, 2007; Zarin, 2016; Zarin, 2017)
  - And for results of trials (Zarin, 2019), although reporting incomplete (Nelson, 2023)
- NIH RePORTER
  - Database of all research grants funded by NIH

Aggregations – integrating many resources

- Clinical – growing tendency of publishers to aggregate resources into comprehensive products
  - Up to Date
  - Essential Evidence Plus – Includes InfoPOEMS, “Patient-oriented evidence that matters”
  - Dynamed Plus

Aggregations (cont.)

- Biomedical research: Model organism databases, e.g., Mouse Genome Informatics
  - Combines genomics and related data, bibliographic database, gene references, etc.
- Consumer: MEDLINEplus
  - Integrates a variety of licensed resources and public Web sites