Section 1: Data Indexing and Retrieval

William Hersh, MD
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
School of Medicine
Oregon Health & Science University
Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: http://informaticsprofessor.blogspot.com
Twitter: @williamhersh

References

Anonymous (2010). Search Engine Optimization Starter Guide. Mountain View, CA, Google. http://www.google.com/webmasters/docs/search-engine-optimization-starter-guide.pdf
Anonymous (2012). From Screen to Script: The Doctor's Digital Path to Treatment. New York, NY, Manhattan Research; Google. http://www.thinkwithgoogle.com/insights/library/studies/the-doctors-digital-path-to-treatment/
Anonymous (2012). State of the Internet, 3rd Quarter 2012. Reston, VA, comScore. http://www.comscore.com/Insights/Presentations_and_Whitepapers/2012/State_of_the_Internet_in_Q3_2012
Barrows, R and Traverso, J (2006). Search Considered Integral. ACM Queue, May, 2006. http://acmqueue.com/modules.php?name=Content&pa=showpage&pid=389
Bastian, H, Glasziou, P, et al. (2010). Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Medicine*. 7(9): e1000326. http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.1000326
Brin, S and Page, L (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*. 30: 107-117. http://infolab.stanford.edu/pub/papers/google.pdf
Coletti, MH and Bleich, HL (2001). Medical subject headings used to search the biomedical literature. *Journal of the American Medical Informatics Association*. 8: 317-323.
Davies, K (2006). Search and Deploy. Bio-IT World, October 16, 2006. http://www.bio-itworld.com/issues/2006/oct/biogen-idec/
Fox, S and Duggan, M (2013). Health Online 2013. Washington, DC, Pew Internet & American Life Project. http://www.pewinternet.org/Reports/2013/Health-online.aspx
Hersh, WR (2009). Information Retrieval: A Health and Biomedical Perspective (3rd Edition). New York, NY, Springer.
Hersh, WR, Weiner, MG, et al. (2013). Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical Care*. 51(Suppl 3): S30-S37.
Insel, TR, Volkow, ND, et al. (2003). Neuroscience networks: data-sharing in an information age. *PLoS Biology*. 1: E17.
Metzger, J and Rhoads, J (2012). Summary of Key Provisions in Final Rule for Stage 2 HITECH Meaningful Use. Falls Church, VA, Computer Sciences Corp. http://assets1.csc.com/health_services/downloads/CSC_Key_Provisions_of_Final_Rule_for_Stage_2.pdf
Purcell, K, Brenner, J, et al. (2012). Search Engine Use 2012. Washington, DC, Pew Internet & American Life Project. http://www.pewinternet.org/Reports/2012/Search-Engine-Use-2012.aspx
Segal, D (2011). The Dirty Little Secrets of Search. New York, NY. New York Times. February 12, 2011. http://www.nytimes.com/2011/02/13/business/13search.html

Section 1: Data Indexing and Retrieval

William Hersh, MD

Professor and Chair

Department of Medical Informatics & Clinical Epidemiology

Oregon Health & Science University

BD2K Open Educational Resources | Oregon Health & Science University

Section 1: Data Indexing and Retrieval

- Information retrieval (IR; aka, search)
  - Big picture
  - Content, indexing, retrieval
- Overview of topics in this section
  - Finding and accessing datasets, indexing, and identifiers
  - Data curation and version control
  - Ontologies
  - Metadata standards
  - Provenance

Introducing myself

William Hersh, MD

Professor and Chair

Department of Medical Informatics & Clinical Epidemiology

School of Medicine

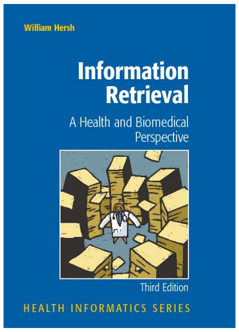
Oregon Health & Science University

Email: hersh@ohsu.edu

Web: www.billhersh.info

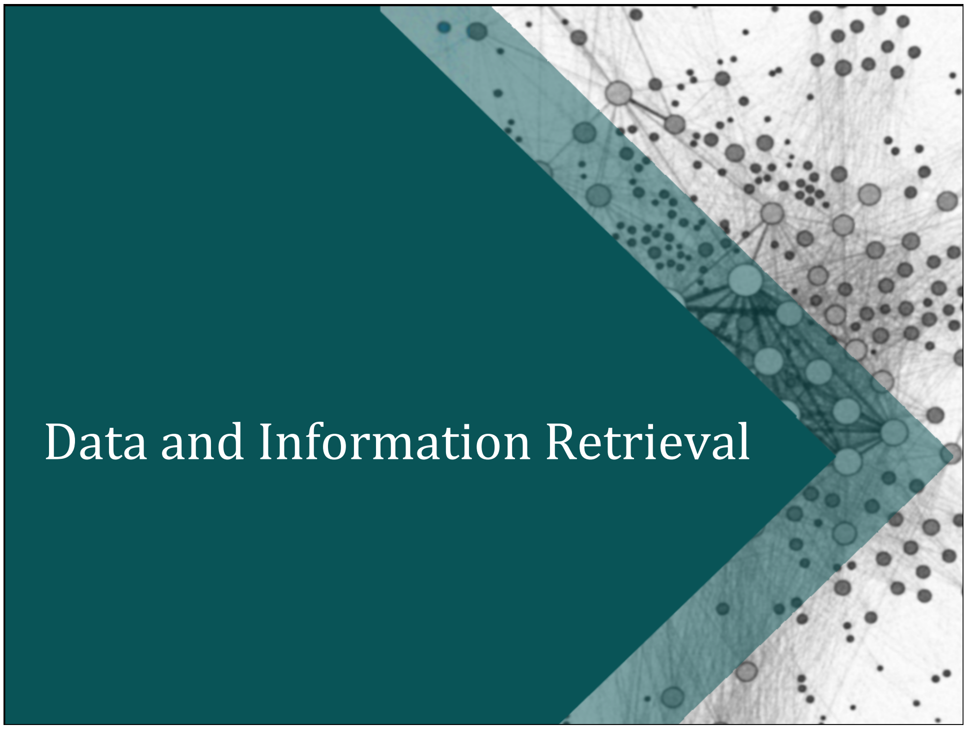
Blog: http://informaticsprofessor.blogspot.com

Twitter: @williamhersh

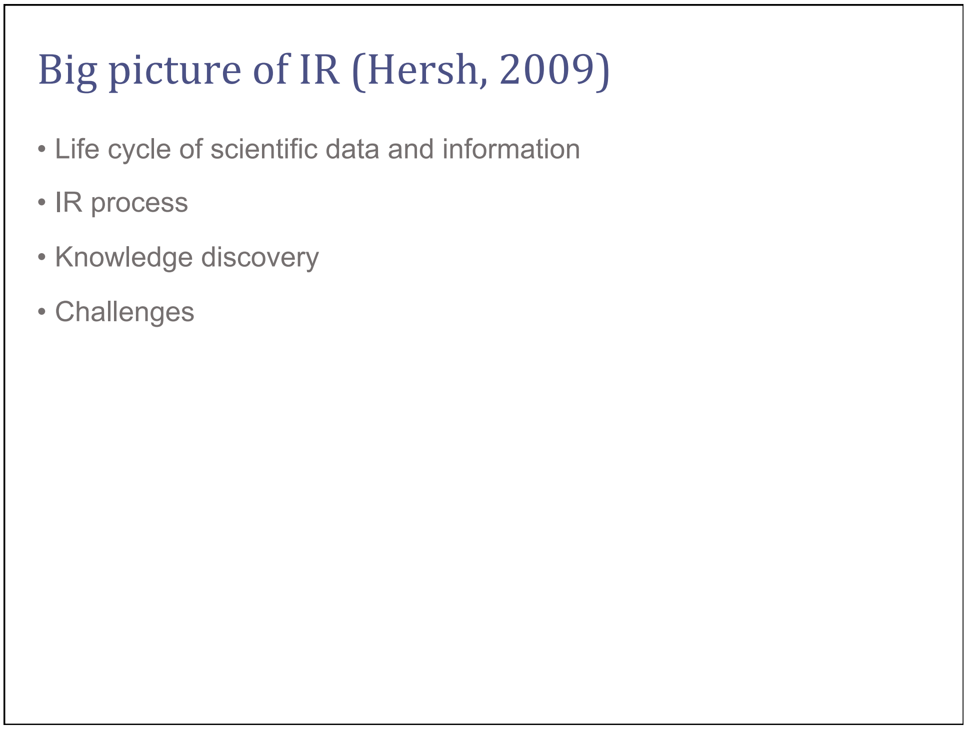

William Hersh
Information Retrieval
A Health and Biomedical Perspective
Third Edition
HEALTH INFORMATICS SERIES

About this lecture

- Content derived from OHSU Open Educational Resources (OERs) content, funded by BD2K Grant 1R25GM114820
  - https://dmice.ohsu.edu/bd2k/
- Will introduce topic with overview of *Information Retrieval* that focuses on organization, indexing, and retrieval of data sources
  - Will also introduce specific topics for rest of Section 1
- Secondary aim is to demonstrate the OERs approach: Taking content from our library to use and re-purpose to teach and learn data science
  - Drawn from BDK10 and BDK14



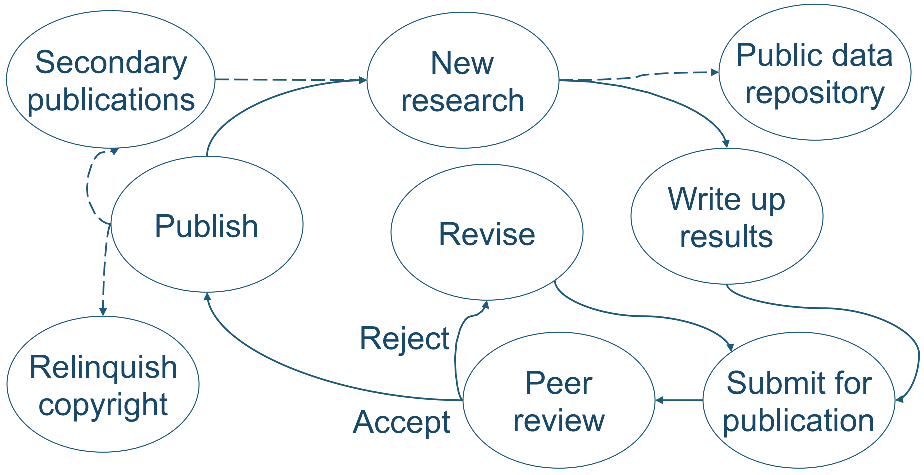
Data and Information Retrieval



Big picture of IR (Hersh, 2009)

- Life cycle of scientific data and information
- IR process
- Knowledge discovery
- Challenges

Life cycle of scientific data and information



Secondary publications

New research

Public data repository

Publish

Revise

Write up results

Relinquish copyright

Reject

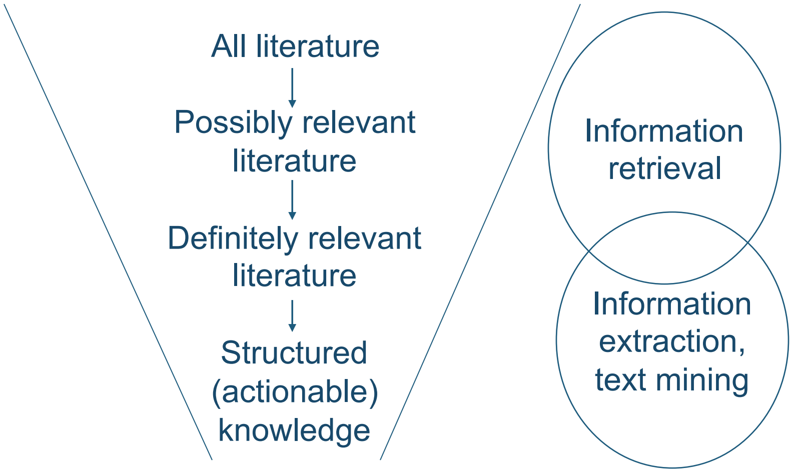
Accept

Peer review

Submit for publication

Notice focus on information and paucity of data (currently)

IR also a growing part of "knowledge discovery" from scientific literature



All literature

Possibly relevant literature

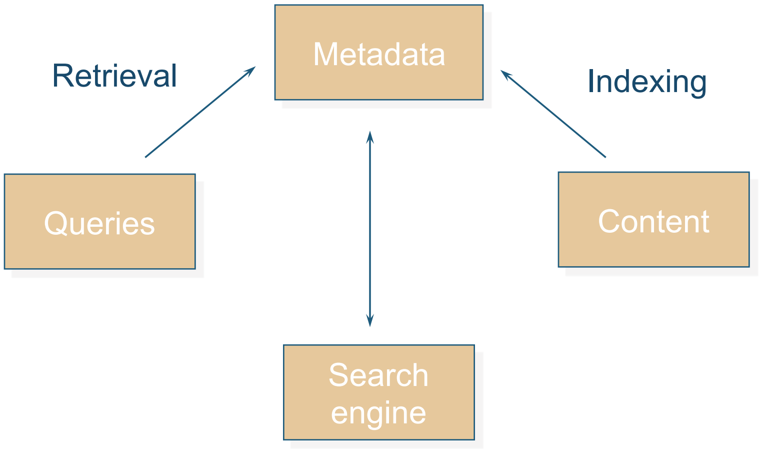
Definitely relevant literature

Structured (actionable) knowledge

Information retrieval

Information extraction, text mining

IR process



Metadata

Retrieval

Indexing

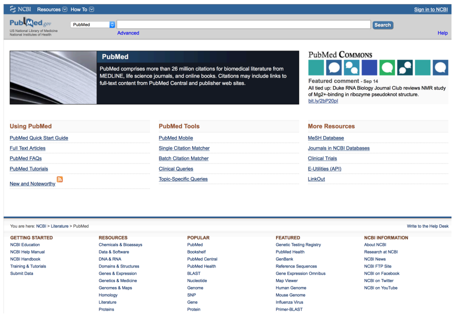
Queries

Content

Search engine

IR is now “mainstream”

- 91% of US Internet users (73% of US adults) have used a search engine
  - 71% of Internet users (59% of US adults) have searched for health information, with 35% using it for self-diagnosis
  - US users now make up only 12% of world Internet population
- “Search” is considered an “integral application”
  - Not only in libraries but also in enterprises and on individual computers and mobile devices
- “Search engine optimization” (SEO) is a key function used by many companies and organizations
  - Many are willing to pay
  - Some are lucky, e.g., last name of “Hersh”



PubMed

PubMed comprises more than 26 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

Featured comment - Sep 14

All tied up: Duke RNA Biology Journal Club reviews NMR study of Mg2+-binding in ribozyme pseudoknot structure.

bit.ly/2bP20pl

Using PubMed

PubMed Quick Start Guide

Full Text Articles

PubMed FAQs

PubMed Tutorials

New and Noteworthy

PubMed Tools

PubMed Mobile

Single Citation Matcher

Batch Citation Matcher

Clinical Queries

Topic-Specific Queries

More Resources

MeSH Database

Journals in NCBI Databases

Clinical Trials

E-Utilities (API)

LinkOut

You are here: NCBI > Literature > PubMed

Write to the Help Desk

GETTING STARTED

NCBI Education

NCBI Help Manual

NCBI Handbook

Training & Tutorials

Submit Data

RESOURCES

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

POPULAR

PubMed

Bookshelf

PubMed Central

PubMed Health

BLAST

Nucleotide

Genome

SNP

Gene

Protein

FEATURED

Genetic Testing Registry

PubMed Health

GenBank

Reference Sequences

Gene Expression Omnibus

Map Viewer

Human Genome

Mouse Genome

Influenza Virus

Primer-BLAST

NCBI INFORMATION

About NCBI

Research at NCBI

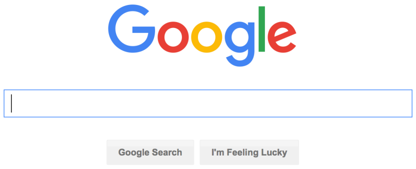
NCBI News

NCBI FTP Site

NCBI on Facebook

NCBI on Twitter

NCBI on YouTube



(Purcell, 2012), (Fox, 2013), (comScore, 2012), (Barrows, 2006), (Google, 2010), (Segal, 2011)

IR and online access firmly planted in health and biomedicine

- Biology is now defined as an “information science”
- Pharmaceutical companies compete for informatics/library talent
- Clinicians cannot keep up – average of 75 clinical trials and 11 systematic reviews published each day
- Search for health information by clinicians, researchers, and patients/ consumers is ubiquitous
  - It’s even part of “meaningful use” – text search over electronic health record notes

(Insel, 2003), (Davies, 2006), (Bastian, 2010), (Purcell, 2012); (Google/Manhattan Research, 2012),(Metzger, 2012)

Some challenges in IR

- We have gone from information paucity to information overload
- Many topics we want to search on have multiple ways to be expressed
  - e.g., diseases, genes, symptoms, etc.
- The converse is a problem too: Many words and terms used to express topics have multiple meanings
  - e.g., cold, Apr-1
- Balancing open access vs. providing for cost of production and maintenance

Content, Indexing, and Retrieval

Content

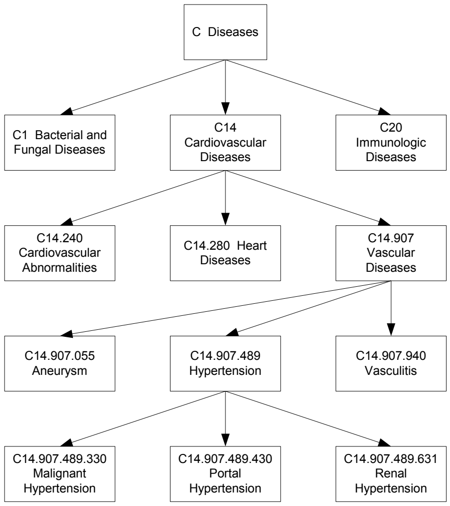
- At first there was bibliographic
  - Abstracts (mostly of journal articles)
- And then (not sequentially timewise),
  - Full-text – articles, textbooks, reports, etc.
  - Hypermedia – including Web pages, images, sounds, videos, etc.
  - Structured collections – textbooks, databases, compendia, etc.
- And now,
  - Data!

Indexing

- Assignment of metadata to content items by automated or manual methods, including
  - Subjects (terms)
    - Automated indexing systems assign words and measures, e.g., frequency
    - Manual indexing systems tend to assign phrases from controlled vocabulary, e.g., MeSH in MEDLINE
  - Attributes
    - Author – e.g., Hersh W
    - Source
    - Publication type – important in evidence-based medicine
    - Secondary source – important for genes, proteins, etc.
    - Grant number
    - Location – link to publisher via DOI

Indexing vocabularies

- Usually
  - Controlled
  - Hierarchical
- Oldest and best known is Medical Subject Headings (MeSH) of NLM
  - Over 26,000 terms, with many synonyms for those terms
  - Hierarchical, based on 16 trees, e.g., Anatomy, Diseases, Chemicals and Drugs
- MeSH browser allows exploration
  - http://www.nlm.nih.gov/mesh/MBrowser.html



C Diseases

C1 Bacterial and Fungal Diseases

C14 Cardiovascular Diseases

C20 Immunologic Diseases

C14.240 Cardiovascular Abnormalities

C14.280 Heart Diseases

C14.907 Vascular Diseases

C14.907.055 Aneurysm

C14.907.489 Hypertension

C14.907.940 Vasculitis

C14.907.489.330 Malignant Hypertension

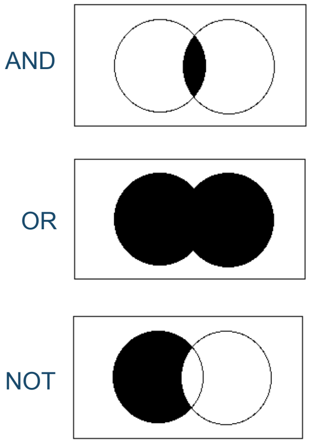
C14.907.489.430 Portal Hypertension

C14.907.489.631 Renal Hypertension

(Coletti, 2001)

Retrieval

- Most common approaches are
  - Boolean – use of AND, OR, NOT
  - Natural language – words common to query and content
- Natural language systems usually rank retrieval results by some measure of relevance
  - Word occurrence and frequencies common to content and query
  - In-links (Google PageRank; Brin, 1998)



Metadata

- "Data about data" – meaning of data elements
- Three types of metadata
  - Descriptive – describes data element for discovery or identification
  - Structural – describes organizational structure of data
  - Administrative – information on how to manage
    - Rights management – who can access and when, how, etc.
    - Preservation – archiving and storage
- In context of IR systems
  - Literature indexing and annotation

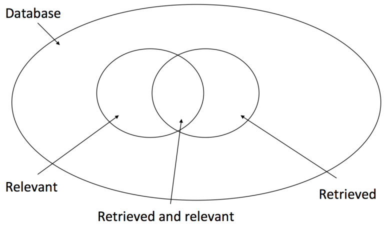
How do we know how well we searched?

- Many questions we could possibly ask (Hersh, 2009)
  - Is system used?
  - Are users satisfied?
  - Do they find relevant information?
  - Do they complete their desired task?
- Most common approach is to measure proportions of relevant documents retrieval
  - Recall

$$R = \frac{\#\text{retrieved and relevant documents}}{\#\text{relevant documents in collection}}$$

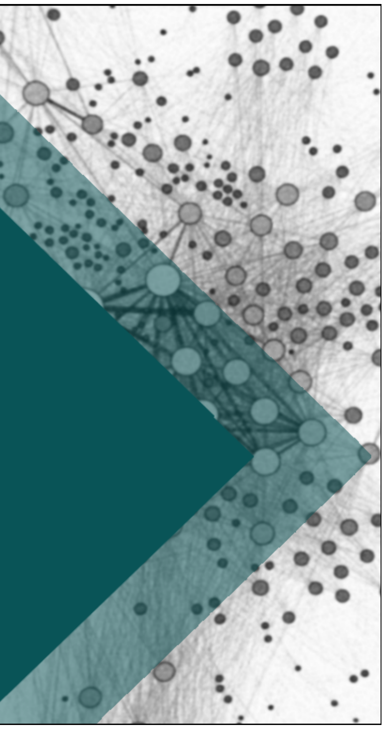
  - Precision

$$P = \frac{\#\text{retrieved and relevant documents}}{\#\text{retrieved documents}}$$



Topics of Section 1

Questions in context of indexing and retrieval

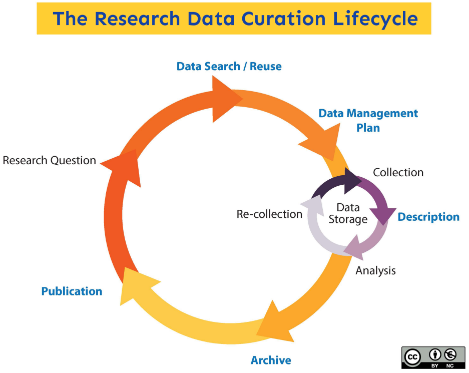


Finding and accessing datasets, indexing, and identifiers

- Many emerging portals indexing and allowing retrieval of data, e.g.,
  - DataCite – www.datacite.org
  - DataMed (from bioCADDIE) – www.datamed.org
- Are there issues unique to indexing and retrieval of data sets?

Data curation and version control

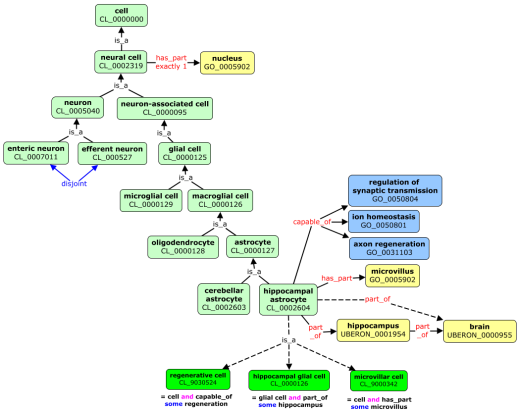
- Data sets have context and may grow or change
- How do we insure data are best described and up to date?
- Data curation and document indexing have overlapping activities



http://library.ucmerced.edu/research/researchers/research-data-curation

Ontologies

- More than a vocabulary or thesaurus, a “formal conceptualization of a specified domain” (BDK14)
  - Terms are defined
  - Relationships between terms are defined, allowing logical inference and sophisticated data queries
  - Terms are arranged in a hierarchy
- How can we optimally index and then retrieval using ontologies?



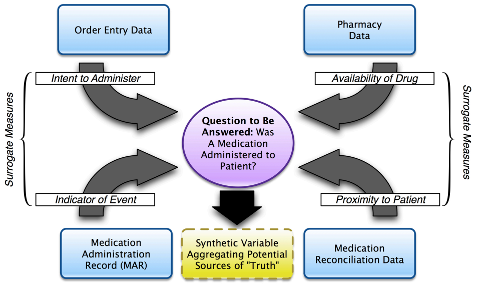
cell CL_0000000
is_a
neural cell CL_0002319
has_part exactly 1
nucleus GO_0005902
is_a
neuron CL_0005040
neuron-associated cell CL_0000095
is_a
is_a
enteric neuron CL_0007011
efferent neuron CL_000527
glial cell CL_0000125
disjoint
is_a
microglial cell CL_0000129
macroglial cell CL_0000126
regulation of synaptic transmission GO_0050804
capable_of
ion homeostasis GO_0050801
is_a
oligodendrocyte CL_0000128
astrocyte CL_0000127
axon regeneration GO_0031103
is_a
has_part
microvillus GO_0005902
cerebellar astrocyte CL_0002603
hippocampal astrocyte CL_0002604
part_of
part _of
hippocampus UBERON_0001954
part _of
brain UBERON_0000955
is_a
regenerative cell CL_9030524
= cell and capable_of some regeneration
hippocampal glial cell CL_0000126
= glial cell and part_of some hippocampus
microvillar cell CL_9000342
= cell and has_part some microvillus

Metadata standards

- Many types, which one(s) to use depends on task
- ISO standards for metadata (generic)
  - http://metadata-standards.org
- Some better-known metadata standards for IR include
  - MEDLINE – focus on biomedical literature; includes MeSH
  - Dublin Core Metadata – motivated by Web resources
  - DataCite – focus on data sources, builds on Dublin Core
  - bioCADDIE – mapping among a variety of metadata standards with a focus on biomedical data
- What approach(es) is/are optimal for data indexing and retrieval?

Provenance

- Where does your data come from?
- Example from clinical medicine (Hersh, 2013): Is a patient on a medication?
- Any re-use of data must take provenance into account
- How do we best describe provenance in data sets?



Order Entry Data

Pharmacy Data

Surrogate Measures

Intent to Administer

Availability of Drug

Question to Be Answered: Was A Medication Administered to Patient?

Indicator of Event

Proximity to Patient

Surrogate Measures

Medication Administration Record (MAR)

Synthetic Variable Aggregating Potential Sources of "Truth"

Medication Reconciliation Data

Questions?

William Hersh, MD

Oregon Health & Science University

hersh@ohsu.edu