

8.3 Indexing

10x10 – Introduction to Biomedical Informatics & Artificial Intelligence
William Hersh, MD, FACMI, FAMIA, FIAHSI
Copyright 2026, All Rights Reserved



Indexing

Assignment of metadata to content to facilitate retrieval

Two major types

- Human indexing with controlled vocabulary
- Automated indexing of all words

World Wide Web indexing

Human indexing

Usually performed by professional indexer with some background in biomedicine

Follows protocol to scan resource and select terms from a controlled vocabulary

Most vocabularies are hierarchical and have specific definitions for when term is to be assigned

Medical Subject Headings (MeSH) (Colletti, 2001)

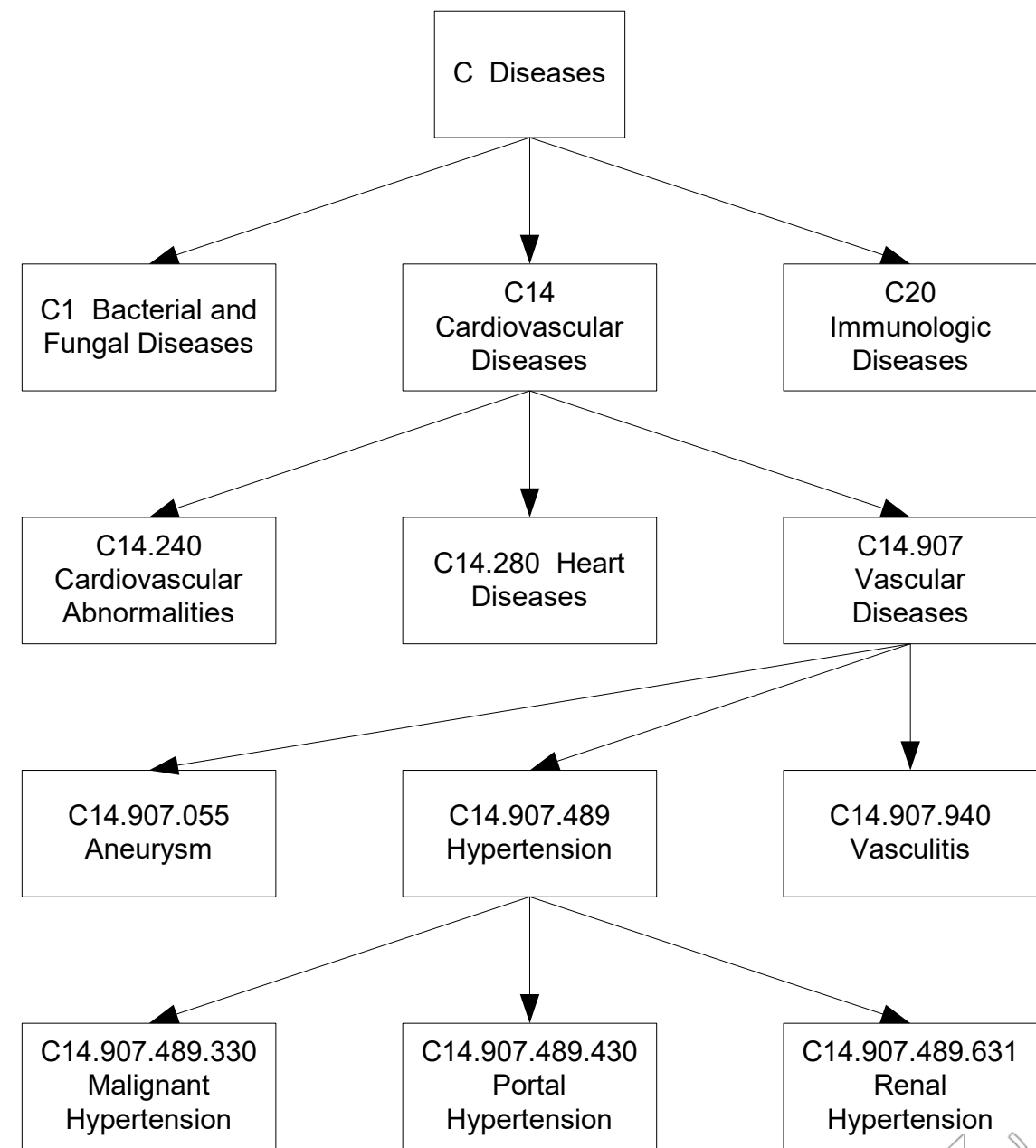
Over 28,000 terms (“headings”), with over 90,000 synonyms (“entry terms”) for those terms

- Supplementary Concept Records are 240,000 additional chemicals, drugs, genes, etc. that indexers encounter in indexing process and map to MeSH headings

Hierarchical, based on 16 trees, e.g., *Anatomy, Diseases, Chemicals and Drugs*

Contains 79 subheadings, which can be used to make a heading more specific, such as *Diagnosis or Therapy*

A slice of MeSH



Many ways to search and browse MeSH

[Search](#)

[Browse by tree](#)

[MeSHOnDemand](#)

[Via PubMed](#)

MEDLINE indexing – then and now

Indexing originally done by professionals who followed protocol first devised by Bachrach (1978)

- Read title, introduction, and conclusion and then scan methods, results, figures, tables, and, lastly, abstract
- Ignore “key words” of publisher
- Assign 2-4 headings (with or without subheadings) as central concepts (or major headings) and another 5-10 as minor headings
- Use most specific headings in hierarchy assigned
- Any publication type assigned, e.g., Randomized Controlled Trial, Meta-Analysis, Practice Guideline, Review

Automated since 2022

Other bibliographic indexing

Other NLM databases use MeSH

Some non-NLM resources use MeSH

- MeSH [freely available from NLM](#)

Other non-NLM databases have their own subject headings, e.g.,

- CINAHL subject headings
- EMTREE

Other metadata

Indexing covers more than content

Other attributes of documents to index can include

- Author(s)
- Source: journal name, issue, pages
- Relationship to other information
 - e.g., gene identifier, grant number, etc.

Automated indexing

Indexing of all words that occur in content items

- In bibliographic databases, will usually include title, abstract, and often other fields, e.g., author or subject heading
- In full-text documents, will usually include all text, including title

Often use a stop word list to remove common words (e.g., the, and, which)

Some systems “stem” words to root form (e.g., coughs or coughing to cough)

Weighted indexing (Salton, 1991)

Often used with automated indexing

Gives weight to words that are frequent but discriminating

Most common approach is for weight to equal product $TF \cdot IDF$

- Inverse document frequency of word i
 - $IDF_i = \log(\# \text{ documents} / \# \text{ documents with word}) + 1$
- Term frequency of word i in document j
 - $TF_{ij} = \text{frequency of word in document}$

Example of a word in different collections, e.g., HIV

- From a collection on HIV
 - The word HIV will likely occur in almost every document, while retinopathy will be much more “discriminating”
- In a general medical collection
 - HIV will occur relatively less frequently, so is better indexing term

Coming full circle – automated indexing of MeSH terms

Explored in much research over the years (e.g., Hersh, 1990) but did not improve retrieval performance much (e.g., Hersh, 1995)

Has been aided in recent years by machine learning approaches (Rae, 2022)

Leading NLM to replace manual MeSH indexing with automated indexing and some manual review (NLM Tech Bulletin, 2021)

“Visual” indexing – e.g., word clouds

Scientific publications of your instructor (from SciVal app)



Citation indexing

Other content items that “cite” this one, e.g., references, links, etc.

Indexing is at content item level

Goal is to designate related or important content items

Citation databases list all other articles that cite a specific article in journals

- e.g., *Science Citation Index*, *SCOPUS*, and *Google Scholar*

Novel feature of Google search engine (Brin, 1998) was giving higher weight to Web pages that have more links to them

Limitations of human indexing

Inconsistency

- When MEDLINE records indexed in duplicate, consistency varies from 63% for central concept headings to 36% for heading-subheading combination (Funk, 1983)
- Results verified even with modern indexing tools and methods (Marcetich, 2004)

Inadequate indexing vocabulary

- Up to 25% of all concepts not represented in MeSH (Hersh, 1994)
- Ambiguities and other naming problems with genes, proteins, etc. (Yandell, 2002; Tuason, 2004)

Limitations of word indexing

Synonymy – e.g., *cancer/carcinoma*

Polysemy – e.g., *lead*

Context – e.g., *high blood pressure*

Focus – e.g., central vs. incidental concepts

Granularity – e.g., antibiotics vs. specific ones

Some other important metadata elements

Digital object identifier (DOI)

- Unique identifier for digital objects
- Standard – ISO 26324
- Used widely, e.g., in MEDLINE/Pubmed and by most journals
- Consists of two parts – which resolve to URL
 - Prefix – publishing entity, e.g., journal or publisher
 - Suffix – assigned by entity, e.g., scientific paper
- My first informatics paper has one (Hersh, 1990)
 - Prefix – 10.1016
 - Suffix – 0010-4809(90)90031-7
 - [URL](#)

Open Researcher and Contributor ID (ORCID)

- Unique identifier for scientific authors
- Useful for author name disambiguation
- You should have one
 - Mine is 0000-0002-4114-5148
 - [URL](#)

Indexing the Web

Both manual and automated approaches

Human indexing

- Sheer volume makes large-scale human indexing unlikely
- Metadata tagging – [Dublin Core Metadata Initiative \(DCMI\)](#)
- Schema.org – metadata terminology for Web (Iliadis, 2023)
- Web catalogs – categorize pages or sites under broad headings

Automated indexing

- Web crawling – following all links and applying word-based indexing