Copyright Clearance Center

RightsDirect
A Copyright Clearance Center Subsidiary

SciBite

WHITEPAPER

The Evolution and Importance of Biomedical Ontologies for Scientific Literature

The volume of scientific literature being published has increased dramatically in the digital age. Ontologies and taxonomies are important tools to help researchers retrieve and understand this overwhelming amount of scientific literature, but using and managing ontologies can be challenging in itself.

In this paper we're looking at the history of biomedical classification and how these systems have evolved to address new technology and use cases. We'll explain the difference between taxonomies and ontologies, and discuss the challenges and successes that come with adopting and managing ontologies.

FROM CLAY TABLETS AND BOOKS TO THE FUTURE

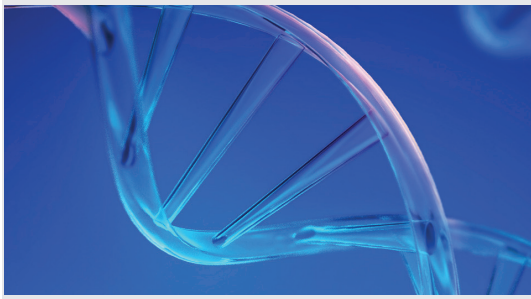
Biomedical classification began with organizing collections in physical space, so people could locate items. Cuneiform tablets, books, specimens, Linnaean taxonomy: subject-matter experts arranged collections to clarify meaning and identify conceptual relationships.

As scientific communication advanced from print publication to electronic indexes, access to relevant information in a timely manner became critical. For biomedical literature, the merging of *Index Medicus* with the *Quarterly Cumulative Index to Current Literature*, published jointly by the AMA and the Army Medical Library in 1927, was a milestone event.[1] That evolved into the National Library of Medicine's Subject Heading Authority list in 1954, published alongside *Index Medicus (New Series)* as the Medical Subject Headings (MeSH) in 1960.[2]

Researchers are now involved in both developing the vocabularies to represent entities and relations within different domains, and in annotating their experimental data to make it available to others.

The start of the 21st century brought a new paradigm of biomedical classification. There has been a revolution in biology from a largely qualitative discipline to a quantitative one. New subdisciplines like bioinformatics and computational biology have emerged to cope with this transition and biologists have had to adapt to this change. It began with the Gene Ontology (GO) project — an effort to classify gene and protein function with a standardized terminology — but soon lead to an explosion of biomedical ontologies in areas such as cells, anatomy, disease, genetics and assays. The linking of primary and clinical data to structured vocabularies, available to be used in algorithmic processing, makes vast amounts of significant data actionable in powerful ways. Researchers are now involved in both developing the vocabularies to represent entities and relations within different domains, and in annotating their experimental data to make it available to others.

As these structured domain resources proliferated and matured, they soon met challenges, such as overlap of scope between or among ontologies, unclear licensing terms, ontologies that are no longer maintained, and the use of undefined relationships between ontology classes. In response, the community formed the OBO Foundry — a consortium of ontology developers who agreed to meet a set of defined standards for their ontologies, such as openness, orthogonality (no term can be a member of more than one group) and use of a common format and set of relationships. Major areas of overlap still exist in some areas of the biomedical ontology realm, especially in the areas of disease and laboratory processes. Efforts are still underway to tackle these challenges. For example, the Mondo Disease Ontology is an effort to align and harmonize the multiple ontologies and terminologies that exist for describing diseases across the world.[3]



THE HISTORY BEHIND MEDICAL SUBJECT HEADINGS (MeSH)

The National Library of Medicine's MeSH was first published in 1960.[4] The next noteworthy leap came in 1964 when Winifred Sewell adapted MeSH for automated indexing in MEDLARS.[5] It is this curated, controlled, hierarchical thesaurus that powers PubMed, Medline, and biomedical searching around the world.

Every descriptor term in MeSH is assigned one or more Concept(s).[6] For each descriptor, there is a Preferred Concept, which includes preferred and synonymous terms, and often also a Narrower Concept. These Concepts are not used in Medline searching or indexing. Concepts are used in creating and updating MeSH terms, and in providing avenues for relationships to other systems, such as the UMLS (Unified Medical Language System).

Anatomy of a MeSH Descriptor

A descriptor in MeSH consists of several parts:

Dementia MeSH Descriptor Data 2020

Details | Qualifiers | MeSH Tree Structures | Concepts

MeSH Heading	Dementia
Tree Number(s)	C10.228.140.380 F03.615.400
Unique ID	D003704
RDF Unique Identifier	http://id.nlm.nih.gov/mesh/D003704
Scope Note	An acquired organic mental disorder with loss of intellectual abilities of sufficient severity to interfere with social or occupational functioning. The dysfunction is multifaceted and involves memory, behavior, personality, judgment, attention, spatial relations, language, abstract thought, and other executive functions. The intellectual decline is usually progressive, and initially spares the level of consciousness.
Entry Term(s)	Amentia Familial Dementia Senile Paranoid Dementia
See Also	Tauopathies
Public MeSH Note	1981; see DEMENTIA, SENILE 1979-91; see PSYCHOSES, SENILE 1963-76; see PSYCHOSES, PRESENILE 1963-72; see PICK DISEASE OF BRAIN 1981-91; see PICK'S DISEASE OF BRAIN 1966-80; for DEMENTIA, PRESENILE see DEMENTIA 1967-97; for DEMENTIA, SENILE see DEMENTIA 1966-97
History Note	1981(1963)
Date Established	1981/01/01
Date of Entry	1980/05/21
Revision Date	2015/06/30

page delivered in 0.117s

The MeSH Heading is what we think of as the main term. Tree Numbers specify where the term fits within the hierarchy. Unique ID belongs to the term regardless of its tree position. Scope Note defines how this term is applied to citations. Entry Terms, also known as "see references" are synonyms and alternate terms mapped to the descriptor. See also include suggested related descriptor terms. Consider also (not shown) includes descriptors related by linguistic roots.

MeSH descriptors are organized into 16 categories. Each category has a hierarchical structure, leading to these being called "trees."[7] A single term can (and often does) occur in more than one tree. Note below the term Dementia is situated with both broader and narrower terms:

Mental Disorders [F03]
- Neurocognitive Disorders [F03.615]
  - Amnesia [F03.615.200]
  - Cognition Disorders [F03.615.250]
  - Consciousness Disorders [F03.615.300]
  - Delirium [F03.615.350]
  - **Dementia [F03.615.400]**
    - AIDS Dementia Complex [F03.615.400.050]
    - Alzheimer Disease [F03.615.400.100]
    - Aphasia, Primary Progressive [F03.615.400.125]
    - Creutzfeldt-Jakob Syndrome [F03.615.400.300]
    - Dementia, Vascular [F03.615.400.350]
    - Diffuse Neurofibrillary Tangles with Calcification [F03.615.400.370]
    - Frontotemporal Lobar Degeneration [F03.615.400.380]
    - Huntington Disease [F03.615.400.390]
    - Kluver-Bucy Syndrome [F03.615.400.431]
    - Lewy Body Disease [F03.615.400.512]
  - Dyslexia, Acquired [F03.615.700]

The MeSH descriptor set has been greatly expanded over time. In 1960 there were 4,400 descriptors, while in 2015 there were 27,455 descriptors in use. The 2020 MeSH release is expected to include nearly 30,000 descriptors.

While the number of descriptors has increased, the number of Qualifiers, commonly called subheadings, has been reduced. Subheadings in the early versions of MeSH were assigned to specific categories and varied in phrasing between trees. For the 1960 MeSH, a smaller set of standardized subheadings was adopted for use across categories. These subheadings "qualify" what aspect of the descriptor is discussed.

For the researcher, these changes present both challenges and opportunities. For example, in 2000 the term 'Multiple Sclerosis' was modified with the addition of two narrower terms: 'Multiple Sclerosis, Chronic Progressive' and 'Multiple Sclerosis, Relapsing-Remitting'. Using the narrow term 'Multiple Sclerosis, Relapsing-Remitting' will retrieve works which discuss this specific variant of the disease. This query can be further restricted to only records where this is a major descriptor, a main focus of the article.

In order to retrieve records discussing all variants of multiple sclerosis, most implementations of MeSH include an "explode" function. Using "explode" retrieves the term entered, as well as all narrower terms in the MeSH tree.

Multiple Sclerosis [C20.111.258.250.500] ⊖
Multiple Sclerosis, Chronic Progressive [C20.111.258.250.500.200]
Multiple Sclerosis, Relapsing-Remitting [C20.111.258.250.500.600]

BEYOND MeSH

While MeSH is highly regarded within the bio- and health sciences world, there are other vocabularies that are also highly valued and heavily used (Figure 1). For example, Emtree, a commercial ontology developed by Elsevier, is a similar hierarchical thesaurus used to search the database Embase. SNOMED, the Systematized Nomenclature of Medicine, is an international standard language for clinical terms. ICD, International Statistical Classification of Diseases and Related Health Problems, is maintained by the World Health Organization. These tools have overlapping concepts and contribute individual strengths.

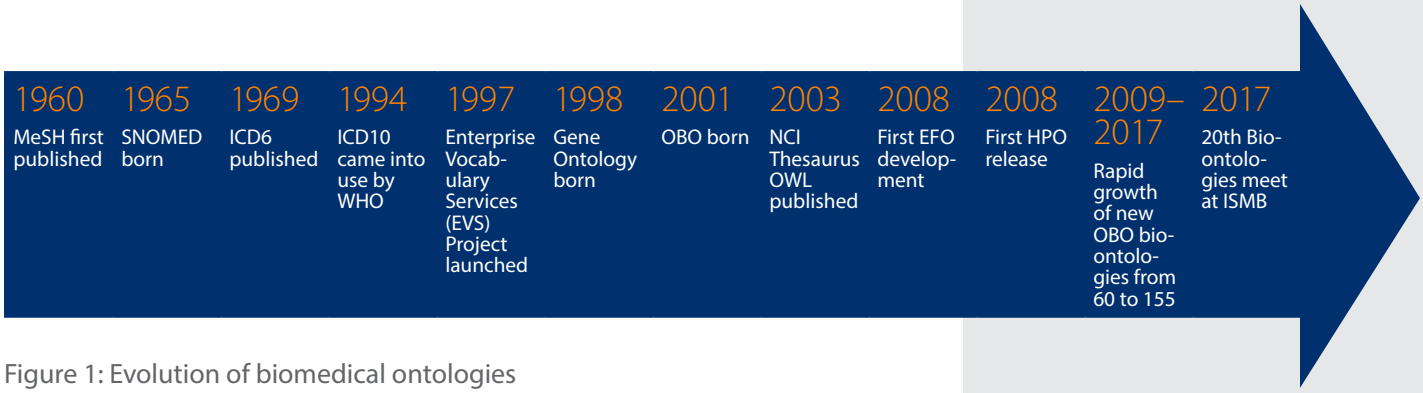


Figure 1: Evolution of biomedical ontologies

Life sciences professionals today have the benefit of this rich set of mature ontologies and terminologies. They provide a key role in the ongoing challenge of managing big data by:

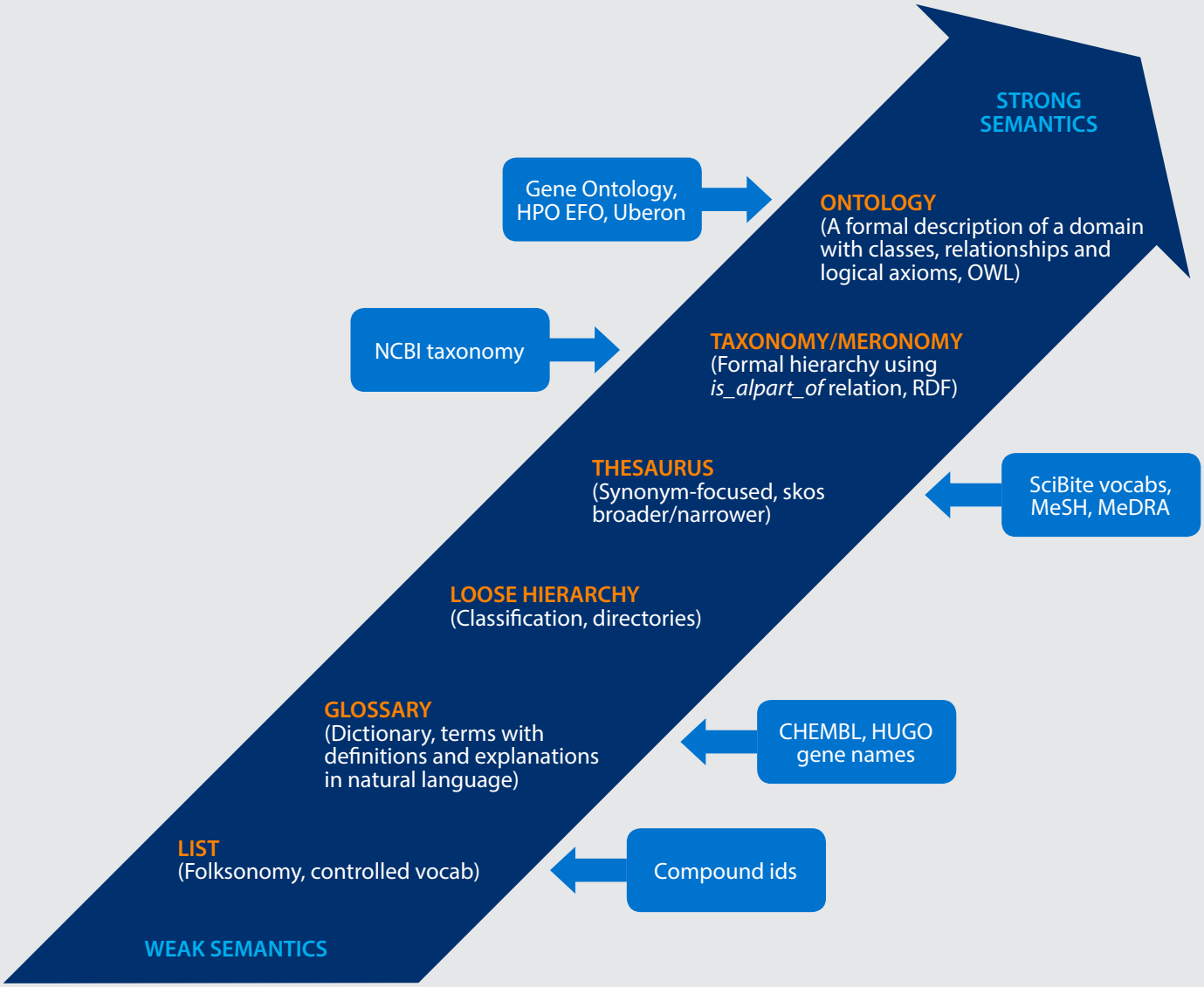
- supporting consistent metadata descriptions across disparate research areas
- encapsulating the human understanding of the life sciences in a way that can be understood by both humans and machines
- providing training data for machine learning and AI

Biomedical ontologies and terminologies are also an essential requirement in the quest for FAIR research data: data that is Findable, Accessible, Interoperable and Reusable. The lack of FAIR data is thought to cost the EU alone at least 10.2 billion euros a year.[8]

Ontology vs Taxonomy

What is the difference between a taxonomy and an ontology?

Ontologies and taxonomies are in many ways similar — they describe types of things, and are arranged in a hierarchical structure. The way they differ is in their level of formality of their semantic relationships: taxonomies tend to relate classes using only is_a, e.g. kidney is_a organ is_a anatomical structure whereas ontologies use a more expressive set of pre-defined relationships such as part_of, develops_from. In fact, both systems of description might be thought of as being on a spectrum with increasingly strong semantics, from a collection of terms (tags) to enhance categorization (a 'folksonomy'), through to an ontology which is a formal logical description of a domain with classes and relationships that operate between them (figure 2, below).



CHALLENGES WHEN USING AND MANAGING ONTOLOGIES

Ontologies can now be used for many applications, including search, data analysis, indexing and information sharing.

The use of ontologies, potentially across structured and unstructured content, across a range of domains, opens possibilities far beyond Boolean searching of pooled citations. For example:

- Synonyms contributed by multiple vocabularies can be added to queries automatically, expanding retrieval while maintaining relevance.
- Synonyms and standardized or related terms suggested to a user can decrease abandoned queries and increase the length and activity of search sessions. It is this synonym support and expansion that enables the structuring of data, and makes ontologies more fit for text mining and analytics.
- Extending managed vocabularies to unstructured text increases the efficiency and reach of pharmacovigilance queries.
- The relationships embedded in advanced ontologies allow users to mine for unexpected co-occurrences and suggest novel uses for existing drugs, or similarities between diseases or functions that might point to underlying processes.

However, adopting and managing ontologies to achieve these benefits can be a challenge. To make use of public ontologies and taxonomies - such as MeSH, MedDRA, SNOMED, and NCI Thesaurus - in named entity recognition scenarios, for example, these must be tailored for use through expert manual curation or automated processes. These and private ontologies and taxonomies require high-throughput software to convert unstructured text into machine-readable information annotated with ontology terms.

Managing ontologies across an organization can also be a challenge. For example, if you use public ontologies modified locally, how do you then reconcile changes made in the original ontology with your modifications? How do you allow subject matter experts and users to contribute changes to your ontology safely and quickly?

Different types of ontology users need different permissions to suggest changes or make changes directly and immediately. Versioning, local changes, and audit trails must also be accounted for in productionized operations.

SUMMARY

The volume of scientific literature has exploded. The value of searching across genetic, clinical, anatomy, and biochemical resources at once has shaped the new paradigm for research and development. Biomedical ontologies are a necessary tool to harness big data so that both humans and machines can advance our scientific understanding.

Jane Lomax

Head of Ontologies, SciBite

Jane leads the development of SciBite's vocabularies and ontology services. She holds a PhD in Genetics from Cambridge University and has 15 years' experience working with biomedical ontologies, including at the European Bioinformatics Institute and Wellcome Sanger Institute. She has published over 35 scientific papers, mainly in the area of ontology development and biocuration, and contributes to public ontology projects including Pistoia Ontology Mapping Project, the OBO Foundry and is a current member of the Executive Committee for the International Society of Biocuration.

Elizabeth Wolf

Elizabeth S. Wolf is Data Quality Manager at Copyright Clearance Center. She earned her MLS at University of Maryland and studied health science reference under Winifred Sewell. Elizabeth is a member of the team responsible for the CCC Managed Data-Works Management System. She also provides User Acceptance Testing (UAT) for RightFind Navigate, an aggregated search platform enhanced by machine learning and contextualized discovery. Elizabeth leads the Expert Literature Search Service, including pharmacovigilance searching. She is a member of Metadata 2020, has served on two NISO working groups, and has extensive experience with ontologies.

[1] Greenberg SJ, Gallagher PE. The great contribution: Index Medicus, Index-Catalogue, and IndexCat. J Med Libr Assoc. 2009;97(2):108–113. doi:10.3163/1536-5050.97.2.007

[2] History of MeSH, accessed at https://www.nlm.nih.gov/mesh/intro_preface.html#pref_hist

[3] Smith B, Ashburner M, Rosse C, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25(11):1251–1255. doi:10.1038/nbt1346 http://obofoundry.org/ontology/mondo.html

[4] History of MeSH, accessed at https://www.nlm.nih.gov/mesh/intro_preface.html#pref_hist

[5] Sewell W. Medical Subject Headings in MEDLARS. Bull Med Libr Assoc. 1964;52(1):164–170.

[6] Concept Structure in MeSH, accessed at https://www.nlm.nih.gov/mesh/concept_structure.html

[7] MeSH Tree Structures, accessed at https://www.nlm.nih.gov/mesh/intro_trees.html

[8] https://op.europa.eu/en/publication-detail/-/publication/d3766478-1a09-11e9-8d04-01aa75ed71a1/language-en

LEARN MORE

RightFind® Insight, powered by the SciBite® platform, brings the power of semantic enrichment to the search and reading experience to turn information into knowledge and accelerate new discoveries. To learn more, visit:

www.copyright.com/business/rightfind-insight/

solutions@copyright.com

+1.978.750.8400 (option 3)