Annotation scheme for PubMed query large set

Aurélie Névéol, February 2009

Definitions and examples for each annotation scheme category

The annotation scheme was derived in part from the UMLS Semantic Groups, described in [1] and [2]. Terms in bold are UMLS Semantic Groups. Terms in Italic are UMLS Semantic Types. Below are definitions and examples for each category in the scheme.

Body Part

UMLS definition: A collection of cells and tissues which are localized to a specific area or combine and carry out one or more specialized functions of an organism. This ranges from gross structures to small components of complex organs. These structures are relatively localized in comparison to tissues.
Examples: Small intestine; index finger; Abducent Nerve

Cell or *Cell Component*

UMLS definition: The fundamental structural and functional unit of living organisms; a part of a cell or the intercellular matrix, generally visible by light microscopy.
Examples: Cancer Stem Cells; T-Cells; Plasma Membrane; Axon

Tissue

UMLS definition: An aggregation of similarly specialized cells and the associated intercellular substance. Tissues are relatively non-localized in comparison to body parts, organs or organ components.
Examples: Abdominal Muscle; Adipose tissue; Subcutaneous Tissue

Chemicals & Drugs

Definition: Includes the following UMLS semantic types: *Antibiotic*; *Biomedical or Dental Material; Carbohydrates; Chemical; Chemical Viewed Functionally; Chemical Viewed Structurally; Clinical Drug; Hazardous or Poisonous Substance; Inorganic Chemical; Pharmacological Substance*; *Vitamin*.
Caveat: Please note that some *Biologically Active Substance*s including *Enzyme*s, *Immunologic Factor*s and *Receptor*s may be better categorized as "Genes and Proteins". Any substance listed in Entrez Gene may be categorized as "Genes and Proteins"
Examples: Acids; Pharmaceutical Adjuvants; 5 beta-Dihydrotestosterone; Hypoglycemic Agent; Xigris; Aspirin
Contrasted Examples: As a *Carbohydrate*, "2,3-Diketogulonic Acid" should be categorized as "Chemicals and Drugs". However, the *Carbohydrate* "AGAR", which is also an *Indicator, Reagent, or Diagnostic Aid* listed in Entrez Gene should be categorized as "Genes, Proteins & Molecular Sequences"

Devices

UMLS definition: A manufactured object used primarily in the diagnosis, treatment, or prevention of physiologic or anatomic disorders; A manufactured object used primarily in carrying out scientific research or experimentation.

Examples: Adhesive Bandage; Insulin Syringe; Tissue Microarray; Euglycemic Clamp

Disorders

Definition: Includes the following UMLS semantic types: *Acquired Abnormality*; *Anatomical Abnormality*; *Cell or Molecular Dysfunction*; *Congenital Abnormality*; *Disease or Syndrome*; *Experimental Model of Disease; Injury or Poisoning; Mental or Behavioral Dysfunction; Neoplastic Process; Sign or Symptom.*

Examples: Diabetes, Ischemic Heart Attack, Alcoholism, Ankle Fracture, etc.

Genes and Proteins

Definition: A gene is defined as the section of DNA that represents the blueprint for the construction of a protein. Usually, a gene and the protein it encodes are referred to by similar names. This category also includes the following semantic types: *Amino Acid, Peptide or Protein*; *Enzyme, Lipid*; *Immunologic Factor*; *Indicator, Reagent, or Diagnostic Aid*; *Gene or Genome*; *Nucleic Acid, Nucleoside or Nucleotide*; *Receptor*. To summarize, any substance listed in Entrez Gene may be categorized as "Genes and Proteins"

Examples: PTX1; POLYSERASE 3; DUARTE BRAIN-SPECIFIC PROTEIN

Living Beings

Definition: Includes the following UMLS semantic types: *Alga*; *Amphibian; Animal; Archeon; Bacterium; Bird; Family Group; Fish; Fungus; Human; Invertebrate; Mammal; Organism*; *Patient or Disabled Group; Plant; Popolation Group; Professional or Occupation Group; Reptile; Rickettsia or Chlamidia;Vertebrate; Virus*

Examples: C57BL Mice; Leishmania; Male; Alfalfa

Research Procedures

Definition: An activity carried out as part of research or experimentation. It includes the following semantic types: *Educational Activity; Molecular Biology Research Technique; Research Activity*.

Examples: Epidemiologic Studies; Real-time PCR; Bibliometric Analysis;

Medical Procedures

Definition: An activity relating to the practice of medicine or involving the care of patients, including diagnosis or treatment procedures, techniques or methods. It includes the following semantic types: *Diagnostic Procedures; Health Care Activity; Laboratory Procedure; Therapeutic or Preventive Procedure*.

Examples: Routine Admission Test; Pap Smear; Appendectomy;

Biological Process or Function

Definition: A process or state which occurs naturally or as a result of an activity. It includes the following semantic types: *Biologic Function; Cell Function; Genetic Function*; *Molecular Function; Natural Phenomenon or Process; Organ or Tissue Function; Organism Function; Physiologic Function*

Examples: Apoptosis; Response Inhibition; Pharmacokinetic interaction

MEDLINE Title

Definition: Title of a paper currently referenced in MEDLINE
Examples: Potential reno-protective effects of a gluten-free diet in type 1 diabetes; Spinal meningioma: relationship between histological subtypes and surgical outcome

Author name

Definition: Name of authors currently publishing in MEDLINE referenced journals or elsewhere
Examples: rockman cb; tahara nobuhiro; christou

Journal name

Definition: Name of a publication currently referenced in MEDLINE
Examples: nejm; BMC BioInformatics; j neurol sci

Navigational Information (other than title, journal and author)

Definition: Including PMID, publication year, dates, volume or issue numbers, page numbers but excluding PubMed search history. Limited text may be included in this category, such as months, volume/page/issue abbreviations, etc.
Examples: "19218484"; "pp 124-56"; "2008"; "2009 Feb;25(2):251-258".

Abbreviations

Webster definition: a shortened form of a written word or phrase used in place of the whole
Examples: EEG; TB; AIDS, DNA

Annotation Guidelines

1. **Check before you annotate**: if you are not sure which category should be assigned to a term you are annotating, make the appropriate verification using the UMLS Knowledge Source Server, MEDLINE, etc.
2. **Err on the side of caution**: if an annotation seems questionable, discard to ensure high quality annotations
3. **Be specific**: in the event of overlapping strings, choose the most specific (eg. "ankle fracture" vs. "ankle")
4. **Be comprehensive**: in the event both a long form and short form of a concept appear in a query, annotate both long and short forms separately (eg. "diabetes mellitus" and "DM" in `diabetes mellitus (DM) treatment and diagnosis`)
5. **Be precise**: in queries including tags, only select the string that corresponds to the concept you are annotating, not the tag (eg. select "smith l" vs. "smith l [au]"). Similarly, in the event of a speculative statement, only the concept part of the query should be annotated (eg. "diabetes" in query `possible diabetes`)
6. **Consider context**: in some cases, a concept might belong to two categories. For example, "salmonella" may be used to refer to the bacteria (category **Living Beings**) or to the infection caused by the bacteria (category **Disorders**). Try to pick the category that seems most likely intended by the user based on the query (eg. in query `salmonella treatment` the disorder seems a better fit than the bacteria.
7. **Multiple annotations**: in some cases, multiple annotations may be inferred from a single string. However, only the most specific concept that is the focus of the query should be annotated:

a. Eg. the query `Spinal meningioma: relationship between histological subtypes and surgical outcome` is the title of an article referenced in MEDLINE. The entire query should be annotated as a "title" but the string "spinal meningioma" should *not* be annotated as a disorder.
b. Eg. in the query `breast cancer mammography`, "breast cancer" should be annotated as a disorder but "breast" should *not* be annotated as a body part. In addition, "mammography" should be annotated as a procedure.

8. **String overlap**: in some cases, one string may contribute to two concepts. For example, in the query `hepatitis A and C diagnosis` "hepatitis" is in fact distributed between two concepts. In this case, both "hepatitis a" and "hepatitis c" should be annotated.
9. **Misspellings**: in some cases, misspellings will appear in the queries. When the concept intended by the user can be reasonably inferred, the annotation should be made (eg. select "cirhosis" as a disorder, even though "cirrhosis" is the correct spelling).
10. **Abbreviations**: Abbreviations are the only circumstance where multiple annotations are permissible. For example, in the query `nejm`, "nejm" should be annotated <u>both</u> as a "journal name" and an "abbreviation".

Sample query annotations according to above scheme and guidelines

<u>Query</u>: `Diabetes Metab Res Rev. 2009 Feb 13;25(2):112-126`
<u>Annotations</u>: "Diabetes Metab Res Rev." as BOTH "journal name" and "Abbreviation"
"2009 Feb 13;25(2):112-126" as "Navigational Information"
"Feb" as "Abbreviation"

<u>Query</u>: `"Blood"[Jour] AND 103[volume] AND 1495[page] AND 2004[pdat]`
<u>Annotations</u>: "Blood" as "journal name"
"103", "1495" and "2004" as "Navigational Information"

<u>Query</u>: `adese albopictus RNAi`
<u>Annotations</u>: "adese albopictus" as "living being"
"RNAi" as BOTH "Biological Process or Function" and "Abbreviation"

<u>Query</u>: `IgA nephritis plasmapheresis`
<u>Annotations</u>: "IgA" as BOTH "Genes and Proteins" and "Abbreviation"
"nephritis" as "Disorder"
"plasmaspheresis" as "Medical Procedure"

References

1. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Medinfo 2001;10(Pt 1):216-20.
2. Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. Journal of Biomedical Informatics 2003;36(6):414-432.