***NLM-Chem Chemical Annotation Guidelines***

This document was developed by:

Rezarta Islamaj, Robert Leaman, and corpus annotators: David Cissel, Cathleen Coss, Carol Fisher, Rob Guzman, Preeti Gokal Kochar, Stella Koppel, Dorothy Trinh, Keiko Sekiya, Janice Ward, Deborah Whitman

National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894, USA

Introduction

The purpose of the project is to create a collection of full text documents (corpus), annotated with the location and identity of interesting chemical entities. The NLM-Chem corpus will be used to improve the quality of the automated taggers distributed by the NCBI BioNLP group and will also be described and distributed as a standalone / independent product. The corpus and tagger improvements are intended to be useful for multiple tasks. To make the task concrete, however, improving retrieval of chemicals is used as a representative downstream task.

The corpus development, technical details and quality measurements are described in:

**NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature,** *Scientific Data*

About this Document

This document is broken into the following sections:

* General Guidelines for Annotation of Chemicals: describes the main ideas and our steps for annotating chemicals
* What to Annotate as Chemicals: contains a listing of examples that help illustrate what should be annotated as chemicals
* What Not to Annotate: contains a listing of some exceptions and special cases that should not be annotated
* How to Annotate OTHER: describes the rules for annotating OTHER spans, which indicate the presence of an embedded chemical appearing within another biomedical mention, such as mutation, gene or protein.

This document will be continuously refined as the project progresses.

General Guidelines for Annotation of Chemicals

**Annotation of chemical names in text**

* For this project, we define a chemical as a substance, mixture, or class of substances specific enough to be named and having a defined composition and/or structure
* Most often context in which the chemical is mentioned, and/or usage as a drug is a strong indicator that the mention should be tagged as a chemical.
* The span to be annotated should be the shortest continuous section of text that identifies the chemical
* The goal is to annotate every mention, including multiples, synonyms and abbreviations
* The goal comprises annotation even when the mention is not a primary topic of the text

Examples:

* Thirty patients were randomized into two treatment groups: Group A, oral ketorolac 10 mg plus intramuscular placebo (1 mL saline solution); or Group B, oral placebo (similar tablet to oral ketorolac) plus intramuscular tramadol 50 mg diluted in 1 mL saline solution.
  + Adjectives that are not part of the chemical (“oral”) should not be included
  + Annotate each chemical mention, even if repeated
* ketorolac **→** D020910 (Ketorolac)
* saline solution **→** D077330 (Saline Solution)
* tramadol **→** D014147 (Tramadol)
* In this study synergist triphenyl phosphate (TPP) dramatically increased the toxicity of fenpropathrin.
  + Abbreviations should be annotated along with their expanded text
* triphenyl phosphate (TPP) **→** C005445 (Triphenyl Phosphate)
* fenpropathrin **→** C044267 (Fenpropathrin)
* Amino acid position: Glutamic acid-443 can influence the orientation of active-site residues
  + When the location of a chemical is mentioned, annotate only the chemical, not the position of the amino acid
  + Glutamic acid **→** D018698 (Glutamic Acid)
* beta-galactosidase complexes from several hydrolase families allowed the identification of residue Glu200 as the proton donor
  + Glu **→** D018698 (Glutamic Acid)

**Linking of chemical names to MESH (or Supplementary Concept Record (SCR))**

* Assign the most specific MeSH/SCR ID that can be found within MeSH/SCR vocabularies to describe chemical in question
* Ideally the mention and the assigned MeSH ID will be the same chemical (i.e. exact match)
* If the mention is a chemical that does not have a matching MeSH/SCR ID, but the mention is a member of an identifiable chemical class, then assign the MeSH ID for the chemical class (usually by searching up the MeSH chemical tree)
* If the chemical is a mixture named as multiple chemicals, whose mentions are clearly separate, or are combined with a dash, annotate separate chemicals by assigning the MeSH ID for each ingredient that comprises that mixture
* A mixture name that lists its components may be annotated with a single MeSH ID/SCR if MeSH/SCR contains an entry for that specific combination
* If multiple chemicals are listed as an overlapping ellipsis (or span), annotate the full phrase as a single mention, and assign MeSH IDs to each chemical in the elliptical string (or span) in order of appearance, separated by semicolons,
* Abbreviations and acronyms for chemicals should be assigned the exact same IDs as the original referent chemical
* A single chemical entity may belong to multiple chemical classes. In this case, annotate multiple chemical class IDs, in the order found in the chemical mention appearing in the article, separated by commas.
* If the appropriate MeSH ID(s) is/are not determinable following the rules presented above--or does/do not otherwise exist in the MeSH/SCR controlled-vocabularies--then input a hyphen (“-”) in the identifier field (for each instance where no identifier is assigned)

**Examples:**

* The text “permethrin” should be annotated as D026023 (Permethrin) since it refers to the specific chemical.
* The text “pyrethroids” and “pyrethrins” should be annotated as D011722 (Pyrethrins) because they both refer to the class of chemicals called Pyrethrins.
* The text “imiprothrin” should be annotated as D011722 (Pyrethrins) because it is a member of the Pyrethrins class and it does not have its own MeSH entry.

**Other examples:**

* Phenol **→** D019800 (Phenol)
* Phenols **→** D010636 (Phenols)
* Pyrethroids **→** D011722 (Pyrethrins)
* Imiprothrin **→** D011722 (Pyrethrins)
* cholesterol **→** D002784 (Cholesterol)
* m-formylstyrene **→** D013343 (Styrenes)
* (l-C16)-DPP-(l-C8)-BTZ **→** D011758 (Pyrroles), C012771 (benzotriazole)
* 1,2-O-isopropylidene-3-N-acryloyl-3-deoxy-5-O-[5-chloro-2-N-(2-methoxy-4-(4-methylpiperazin-1- yl)phenyl)pyrimidin-4-yl]-α-d-ribofuranoside 🡪 D011743 (Pyrimidines), D012266 (Ribose), D006027 (Glycosides)

**Examples from our corpus:**

* Three spectra acquisitions for each talc–PLA blend sample have been carried out, for a total of 39 spectra.
  + “talc” 🡪 D013627 (Talc)
  + “PLA” 🡪 C033616 (polylactic acid)
* The ethyl acetate extract (40 g) was subjected to chromatography on silica gel (60–120 mesh, B.D.H.) column using graded solvent systems of petroleum ether–ethyl acetate.
  + “Silica gel” 🡪 D058428
  + “petroleum ether” 🡪 C004544
  + “ethyl acetate” 🡪 C007650
* Levels of oxylipins (5,6-, 8,9-, 11,12- and 14,15-EET, 8-, and 9-HETE, thromboxane B2 [TxB2] were determined in pre- and post-CRH heart perfusates of WT and t-AUCB-treated WT mice through liquid chromatography, tandem mass spectroscopy (LC-MS/MS) as described previously.
  + “5,6-, 8,9-, 11,12- and 14,15-EET” 🡪 C040776; C050715; C046783; C046782

Semicolon separates each individual chemical, the order is preserved.

* + 5,6-EET is C040776 (5,6-epoxy-8,11,14-eicosatrienoic acid)
  + 8,9-EET is C050715 (8,9-epoxyeicosatrienoic acid)
  + 11,12-EET is C046783 (11,12-epoxy-5,8,14-eicosatrienoic acid)
  + 14,15-EET is C046782 (14,15-epoxy-5,8,11-eicosatrienoic acid)
* “8-, 9-HETE”🡪 C047628; D006893
  + 8-HETE is C047628 (8-hydroxyeicosatetraenoic acid)
  + 9-HETE is D006893 (Hydroxyeicosatetraenoic Acids)
* “thromboxane B2”, “TxB2” 🡪D013929 (Thromboxane B2)
* t-AUCB 🡪 C524106 (4-(4-(3-adamantan-1-ylureido)cyclohexyloxy)benzoic acid)

What to Annotate as Chemicals:

Note that this section describes two distinct things: at first it describes what kinds of mentions should be annotated, then what kinds of concepts should be annotated.

**Annotate chemical common names**

* Brand names
* Trade names
* Trivial or common names

**Annotate drug names, as well as substances that are used as drugs**

* British Approved Name (BAN)
* Brand names
* Generic drug names—ketorolac, tramadol
* International Nonproprietary Name (INN)
* Trade names
* Trivial or common names
* United States Adopted Name (USAN)

**Example:**

* Calculated antibiotic treatment was initiated with amoxicillin/clavulanic acid and metronidazole.
  + amoxicillin/clavulanic acid is a specific drug combination and should be highlighted as the whole span 🡪 D019980 (Amoxicillin-Potassium Clavulanate Combination)
  + metronidazole 🡪 D008795 (Metronidazole)

**Annotate systematic chemical names**

* IUPAC names
* IUPAC-like names

**Examples:**

* 2-Acetoxybenzoic acid 🡪 D001241 (Aspirin)
* 2-Acetoxybenzenecarboxylic acid 🡪 D001241 (Aspirin)
* 3,5,4'-trihydroxy-trans-stilbene 🡪 D000077185 (Resveratrol)
* n-(4-hydroxyphenyl)acetamide 🡪 D000082 (Acetaminophen)

**Examples from our corpus:**

* To visualize lipid membranes, sections were stained with 1,1'-dioctadecyl-3,3,3',3'-tetramethylindocarbocyanine perchlorate ('DiI';DiIC18; Molecular Probes) as previously described.
  + 1,1'-dioctadecyl-3,3,3',3'-tetramethylindocarbocyanine perchlorate 🡪 C024286 (3,3'-dioctadecylindocarbocyanine)
  + Dil 🡪 C024286 (3,3'-dioctadecylindocarbocyanine)
  + DilC18 🡪 C024286 (3,3'-dioctadecylindocarbocyanine)
* Heterologous expression combined with 3-(4,5-dimethyl-thiazol-2-yl)-2,5-diphenyltetra-zolium bromide (MTT) cytotoxicity assay in Spodoptera frugiperda (Sf9) cells revealed that PcE1-, PcE7- or PcE9-expressing cells showed significantly higher cytoprotective capability than parental Sf9 cells against fenpropathrin, demonstrating that PcEs probably detoxify fenpropathrin.
  + 3-(4,5-dimethyl-thiazol-2-yl)-2,5-diphenyltetra-zolium bromide, MTT 🡪 C000598529 (3-(4,5-dimethyl-2-thiazolyl)-2,5-diphenyl-2H-tetrazolium bromide)
  + Fenpropathrin 🡪 C044267 (fenpropathrin)

**Annotate chemical formulas**

* InCHI
* InChIKey
* Molecular formulas
* SMILES
* Structural formulas

**Examples from our corpus:**

* The PTM1 stock solution (Invitrogen 2002) contained 5.0 ml of 69% H2SO4, 3.84 g CuSO4, 0.08 g NaI, 3.0 g MnSO4·H2O, 0.2 g Na2MoO4·2H2O, 0.02 g H3BO3, 0.92 g CoCl2·6H2O, 20.0 g ZnCl2 and 65.0 g FeSO4·7H2O l−1.
* NaI 🡪 D012974 (Sodium Iodide)
* MnSO4·H2O 🡪 C039798 (manganese sulfate)
* Na2MoO4·2H2O 🡪 C024687 (sodium molybdate(VI))
* H3BO3 🡪 D001888 (Boric Acids)
* CoCl2·6H2O 🡪 C018021 (cobaltous chloride) Note: hexahydrate is not an ET
* ZnCl2 🡪 C016837 (zinc chloride)
* FeSO4·7H2O 🡪 C020748 (ferrous sulfate)
* Other examples:
* C9H8O4 🡪 D001241 (Aspirin) OR could be other chemicals: acetozone or caffeic acid
* (CH3)2SO 🡪 C007482 (dimethyl sulfate)
* CH3CH2OH 🡪 D000431 (Ethanol)
* CC(=O)Oc1ccccc1C(=O)O 🡪 D001241 (Aspirin)
* InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12) 🡪 D001241 (Aspirin)

**Annotate chemical database identifiers**

Database identifiers in the text should be annotated if they refer to chemicals

* CAS numbers
* ChEBI IDs
* CHEMBL IDs
* PubChem identifiers

**Examples from our corpus:**

* Triton-X100 (0.1% v/v) (CAS No.: ‎9002-93-1) was used as surfactant in all the solution.

CAS refers to the Chemical Abstracts Service registry, a chemical database

* Triton-X100 🡪 D017830 (Triton X-100)
* 9002-93-1🡪 D017830 (Triton X-100)

**Annotate chemical abbreviations and acronyms**

Annotate abbreviations & acronyms that refer to a chemical in context, even if the same text is used elsewhere to refer to non-chemicals

* + Annotate both common abbreviations & acronyms and ones invented by the authors
  + Exclude arbitrary assignments of chemicals to numbers of letters, such as “compound 3”.

**Examples from our corpus:**

* We hypothesized that inhibition of sEH by t-AUCB enhances CRH in isolated mouse hearts through changing the oxylipin profiles, including an increase in EETs/DHETs ratio.
  + t-AUCB 🡪 C524106 (4-(4-(3-adamantan-1-ylureido)cyclohexyloxy)benzoic acid)
  + oxylipin 🡪 D054883 (Oxylipins)
  + EETs🡪 D015777 (Eicosanoids)
  + DHETs 🡪 D015777 (Eicosanoids)
* In addition, the present protocol could be readily extended to the challenging synthesis of cyclohexenyl and cyclohexyl based trifluoromethyl enone 1p and 1q.
  + Annotate the ellipsis as a single mention with 2 IDs separated by a semicolon, but do not pick up the numbers/letters in this string

**Annotate synthetic polymers and manufactured materials of defined composition**

Note: synthetic long peptides are excluded, unless used as a drug (see What Not to Annotate)

**Examples:**

* Poly(l-lactic acid) 🡪 C033616 (poly(lactide))
* Peptoid compounds 🡪 D034444 (Peptoids)
* Nylon 🡪 D009757 (Nylons)
* Polystyrene 🡪 D011137 (Polystyrenes)
* Polystyrene Sulfonate 🡪 C003321 (polystyrene sulfonic acid)
* Polyvinyl chloride (PVC) 🡪 D011143 (Polyvinyl Chloride)
* Polyamides 🡪 D009757 (Nylons),
* Polyacrylamide (PAM) 🡪 C016679 (polyacrylamide)
* Nafion 🡪 C040402 (perfluorosulfonic acid)

**Annotate laboratory reagents which are named and have defined composition**

Note: Excludes peptide and protein reagents (see What Not to Annotate)

Includes:

* Author-defined reagents defined within the article
* Known dyes and indicators
* Known buffers and culture media
* Named proprietary commercial reagents

Examples:

* + fast blue conjugate dye 🡪 C031455 (Fast Blue)
  + Laemmli buffer 🡪 C088816 (Laemmli Buffer)
  + amplex red reagent 🡪 C470430 (Amplex Red)

**Annotate minerals**

Example:

o Talc is a hydrated magnesium silicate, with the general formula Mg3SiO10(OH)2.

• Talc 🡪 D013627 (Talc)

• magnesium silicate🡪 D017633 (Magnesium Silicates)

• Mg3SiO10(OH)2🡪 D017633 (Magnesium Silicates)

**Annotate chemicals found within technique names**

* SDS-PAGE 🡪 D012967 (Sodium Dodecyl Sulfate)
* 13C-NMR 🡪 C000615229 (Carbon-13)

**Annotate chemical/biochemical families and classes, that are generally categorized in the following parts of the MeSH hierarchy:**

1. Any concept that is listed as a descendent of any of the following concepts
   * Inorganic chemicals
   * Organic chemicals
   * Heterocyclic compounds
   * Polycyclic compounds
   * Polymers
   * Carbohydrates
   * Lipids
   * Amino acids
   * Nucleosides
   * Nucleotides
2. Concepts listed under Protein branch will not be annotated as Chemical, unless they are used as a drug, i.e.:
   * + - “Monoclonal antibodies” are considered chemicals, when the context specifies that they are used as drugs
       - Bispecific antibodies are considered chemicals, when the context specifies that they are used as drugs

**Examples from our corpus:**

For descriptive concepts, MeSH tree needs to be consulted to determine if the concept has the right ancestors. For supplementary concepts, all MeSH trees of all listed MeSH headings are considered.

Note, the following branches are not included:

* Hormones: this is a functional class
* Complex Mixtures: these subclasses are mostly descriptive, such as where they come from (“waste products”) or what form they take (“gels”).
* Biological Factors: this is a functional class similar to hormones
* Biomedical and Dental Materials: this is a class of substances useful because of their physical, not chemical properties
* Pharmaceutical Preparations: This mostly lists functional or high-level concepts (e.g. “dosage forms” “plant extracts”).
* Chemical Actions and Uses: this includes concepts like Pharmacologic Actions and Pesticides

While the above are not used to define chemicals, MeSH allows multiple ancestors and individual concepts are typically listed under both functional and structural classes. For example, the hormone testosterone is also listed as a Polycyclic Compound, the pesticide Permethrin is listed as an Organic Chemical, and Silica Gel is also listed under inorganic chemicals, thus all three are classified as chemicals.

**Examples:**

What Not to Annotate

* Do not annotate Homonyms, since they are not chemicals in context (i.e. “reasonable lead time….”)
* Do not annotate Objects and Equipment (i.e. petri plates, litmus paper, molecular sieves, etc)
* Do not annotate Chemical Bonds (i.e. H-bond, C-H, C=O, O-H, N-H etc.), unless written out in words (hydrogen bonds, carbon-oxygen bonds, etc.), in which case we annotate the chemical.
* Do not annotate proteins and peptides (unless they are being used as drugs and have drug/brand names e.g. therapeutic monoclonal antibodies such as rituximab, nivolumab, and the like…)
* Do not annotate mutations of proteins (and genes) when they are presented in one -letter code format of A#A (or B>B, or B/B, etc. in genes), where A stands for amino acid and B stands for nucleotide base abbreviations for the amino acids (i.e. G for glycine) and the bases (i.e. G for guanine). We capture amino-acids and bases that are written out.
* Do not annotate chemical groups that are part of a larger molecule, if such groups are not readily found in MeSH/SCR vocabulary, such as -CH3, methyl, ethyl, ‘X-chemical’ rings, ‘X-chemical’ moiety, etc.
* Do not annotate sequences given for chemicals (including amino-acid sequences of proteins/peptides and base sequences for genes)
* Do not annotate large biomacromolecules and biocomplexes that include protein components (i.e. organelles, ribosomes, cell membranes)
* Do not annotate complex mixtures that are of vague, ill-defined composition (includes lipid bilayer, glass, steel, cement, paper, crude oil, natural gas, etc)
* Do not annotate chemical functions or pharmaceutical actions (i.e. pesticides, analgesics, contraceptives, hypertensives, enzyme Inhibitors, etc.)
* Do not annotate very general words, like:
  + Atoms
  + Matter
  + Element
  + Molecule
  + Gene
  + Protein
  + DNA (includes cDNA, ssDNA, etc.)
  + Polymer
  + Alkali
  + Solutions
  + RNA (includes dsRNA, lncRNA, mRNA, miRNA, rRNA, siRNA, snRNA, snoRNA, tRNA, etc.)
* Do not annotate water if environmental
* lake water
* river water
* seawater

**Examples from our corpus:**

* In the title compound, C31H23NO2S, the pyrrolidine ring adopts an envelope conformation (with the spiro C atom as the flap), while the thiazolidine ring and the two cyclopentane rings adopt twisted conformations.
  + pyrrolidine, thiazolidine and cyclopentane—all 3 are readily found in MeSH/SCR vocabulary (though part of a whole)
* Interactions with Tyr1230 of the activation loop involve a small variety of arenes: several halogenated or nitrated phenyl groups and a larger number of fused 5- and 6-membered heteroatomic aromatic ring systems, mostly in the same space.
  + halogenated or nitrated phenyl groups & fused 5- and 6-membered heteroatomic aromatic ring systems--ignored (not found in MeSH)
  + Tyr 🡪 D014443 (Tyrosine)
  + arenes 🡪 D006841 (Aromatic Hydrocarbons)
* Meanwhile, the thiazolidine ring and the two cyclopentane rings (S1/N1/C23–C25, C1/C2/C10–C12 & C13–C15/C20/C21) are twisted about C25–S1 bond
  + (S1/N1/C23–C25, C1/C2/C10–C12 & C13–C15/C20/C21)--chemical sequence is ignored
  + C25–S1 bond—chemical bonds are ignored
* The cells were stimulated with either 100 nM NPS-Val(6) (SFRNGVGTGMKKTSFQRAKS) or 100 nM NPS-Leu(6) (SFRNGLGTGMKKTSFQRAKS) (both from Proteogenix, France) for 1, 3, 6, and 24 h.
* Drug design of protein kinase inhibitors is now greatly enabled by thousands of publicly available X-ray structures, extensive ligand binding data, and optimized scaffolds coming off patent.
  + Protein kinase inhibitors is a pharmacological action which is not annotated
* Because of the prolonged suppression of paclitaxel-induced neuropathy after removal of cannabinoid agonists, we chose to analyze transcriptional changes in markers of glial activation.
  + Paclitaxel 🡪 D017239
  + Cannabinoid 🡪 D002186 (chemical is found within a pharmacological action phrase)
* Pyrethroids insecticides, analogues naturally occurring pyrethrins extracted from dried flowers Chrysanthemum cinerariaefolium, have been widely used in the control of mites and pests, contributing to more than 25% of world insecticide sales for their high efficiency, broad-spectrum and relatively low toxicity.
  + Insecticides is not annotated
* However, internalization was unaffected when the cells expressed the W22A mutant, which has a mutation in the N-terminal acidic region that destroys the protein’s ability to bind and activate Arp2/3.
  + W22A—ignored, mutation expressed in one-letter code
* Single nucleotide polymorphisms (SNPs) close to the gain-of-function substitution, Asn(107)Ile (rs324981, A>T), in Neuropeptide S Receptor 1 (NPSR1) have been associated with asthma. Furthermore, a functional SNP (rs4751440, G>C) in Neuropeptide S (NPS) encodes a Val(6)Leu substitution on the mature peptide that results in reduced bioactivity.
* Lysates of HeLa cells infected for different lengths of time with live (A) or heat-killed (B) C. burnetii were analyzed by SDS-PAGE and Western blot using an antibody against phosphoTyr421-cortactin (P-cortactin) or an anti-GAPDH antibody.
  + SDS 🡪 D012967 (Sodium Dodecyl Sulfate)
  + phosphoTyr 🡪 D019000 (Phosphotyrosine)

How to Annotate OTHER

The purpose of OTHER annotations is to identify a chemical text span, which in the given context is part of a larger text span referring to a non-chemical, other entity such as a gene or protein. Abbreviations and synonyms to mentions annotated as OTHER category may be ignored

**Examples from our corpus:** (Other annotations, are highlighted in grey, chemicals are highlighted in yellow)

* A recent study about the citrus red mite revealed that a Phe1538 to Ile mutation from the sodium channel played a crucial role in fenpropathrin resistance after comparison of field fenpropathrin-resistant (WZ) and susceptible strains
  + Phe, Leu—annotate both as CHEMICALS
  + Sodium channel –OTHER, sodium is embedded in a protein
  + Fenpropathrin 🡪 C044267 (fenpropathrin)
* ALP: alkaline phosphatase; TBA: total bile acids; γ-GT: gamma glutamyl transpeptidase; N: normal; LFT: liver function tests; N/A: not available; BRIC 2: benign recurrent intrahepatic cholestasis; sdPSC: small duct primary sclerosing cholangitis.
  + TBA🡪 D001647 (Bile Acids)
  + Bile acids 🡪 D001647 (Bile Acids)
* CMV cytomegalovirus, LDH lactate dehydrogenase, PCP Pneumocystis jirovecii pneumonia, PI3K phosphoinositol-3-kinase.

*Note that the abbreviations of OTHER proteins (enzymes) are ignored/not annotated*

* This inc*r*eases IP3, which then interacts with the IP3 receptor (IP3R) to stimulate release of Ca2+ from intracellular stores.
  + IP3 🡪 D015544 (Inositol 1,4,5-Trisphosphate)
  + Ca2+ 🡪 D002118 (Calcium)
* We show that scaffold assembly requires conserved leucine zipper domain to assemble in vitro.
  + leucine—annotated as CHEMICAL
* Only a few years later, the adoption of tyrosine kinase inhibitors (TKIs) began with the use of imatinib in the treatment of chronic myeloid leukemia (CML).
* Tyrosine kinase—OTHER string, though found within a pharmacological action
* Imatinib 🡪 D000068877 (Imatinib Mesylate)