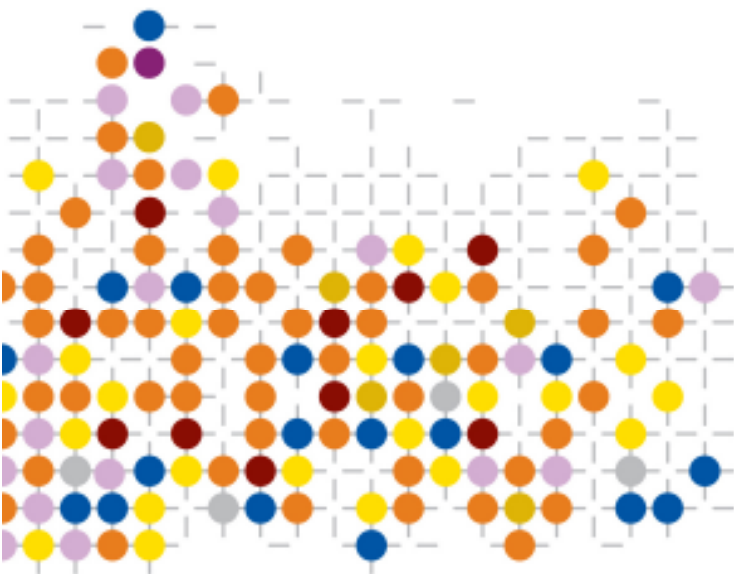


# Emtree

The Life Science Thesaurus  
*2009*



# Table of contents

<b>1. Emtree history</b>	<b>3</b>
<b>2. Emtree content</b>	<b>3</b>
<b>3. Emtree structure</b>	<b>3</b>
<b>4. Unique aspects of Emtree</b>	<b>5</b>
<b>4.1. Natural language terminology</b>	<b>5</b>
<b>4.2. Scope notes</b>	<b>5</b>
<b>4.3. Check tags</b>	<b>6</b>
<b>4.4. Subheadings (previously known as Link terms)</b>	<b>6</b>
<b>4.5. Extensive synonyms</b>	<b>6</b>
<b>4.6. MeSH terms</b>	<b>6</b>
<b>4.7. CAS registry numbers</b>	<b>7</b>
<b>4.8. Enzyme Commission (EC) numbers</b>	<b>7</b>
<b>5. Emtree maintenance</b>	<b>7</b>
<b>5.1. New terms</b>	<b>7</b>
<b>5.2. Changed terms</b>	<b>7</b>
<b>5.3. New structure</b>	<b>8</b>
<b>5.4. Creation dates</b>	<b>8</b>
<b>Appendix 1. Emtree Facts</b>	<b>9</b>

# Emtree: The Life Science Thesaurus

Emtree is Elsevier's Life Science Thesaurus. It is a hierarchically structured, controlled vocabulary, for Biomedicine and related Life Sciences.

Emtree provides a consistent description of biomedical information:

- it offers indexers a comprehensive vocabulary to describe the content of biomedical data
- for database users, it facilitates comprehensive searching and high precision retrieval

## 1. *Emtree history*

Emtree evolved from the Master List of Medical Terms (MALIMET), originally created in 1963 to control and standardize the subject indexes of over 40 printed *Excerpta Medica* abstract journals. But MALIMET grew rapidly as new terms were proposed by the abstract journal editors, until by the end of the 80's it contained an unwieldy 250,000 preferred terms.

Emtree was created in 1988 by pruning MALIMET to the most frequently indexed 25,000 terms, and imposing upon these terms a hierarchical structure based on that of MeSH, the thesaurus used by the U.S. National Library of Medicine for Medline and PubMed. Over the past 21 years, Emtree has been developed and broadened to its current size: a thesaurus comprising over 56,000 preferred terms, which is actively maintained and updated every year (see section 5).

Emtree is integrated into Elsevier's Bibliographic Databases, Embase and Embase.com, and is an important part of the Elsevier Customized Bibliographic Services (EMSCOPES). Emtree is also available in a three-volume print format.

## 2. *Emtree content*

Emtree comprises over 56,000 preferred terms, most of which have one or more synonyms that are available as entry points for indexers and database users. The number of synonyms can be very large: for example the drug isoniazid has more than 200 synonyms, including many trade names and laboratory codes by which it is known.

Overall Emtree has more than 230,000 synonyms: an average of more than four synonyms for each preferred term.

For the user, this means that searching does not require advance knowledge of Emtree's preferred terminology; any phrase describing the required concept is enough. Simply by looking up the phrase in the alphabetical index or (more generally) by checking any of its component words in the permuted index (of all Emtree's words), users can identify the Emtree keywords - preferred terms or synonyms - for the concepts they have in mind.

## 3. *Emtree structure*

Emtree preferred terms are not independent of each other, but are organised into a hierarchical structure with 15 branches, which we call facets:

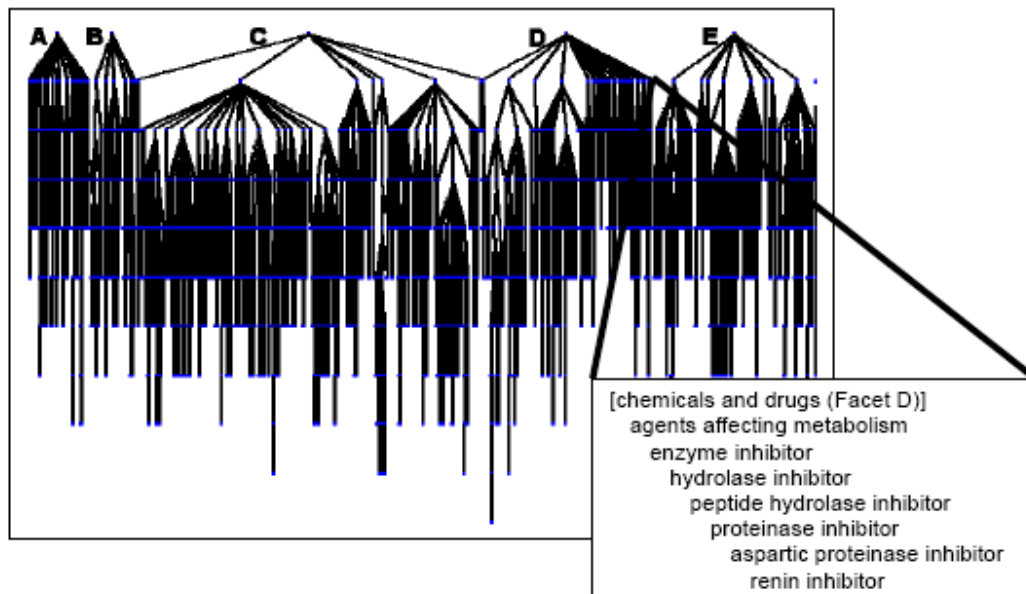
A Anatomical concepts

- B Organism names
- C Physical diseases, disorders and abnormalities
- D Chemicals and drugs
- E Analytical, diagnostic and therapeutic techniques, equipment and parameters
- F Psychological and psychiatric phenomena
- G Biological phenomena and functions
- H Chemical, physical and mathematical phenomena
- I Society and environment
- J Types of article or study
- K Geographic names
- L Groups by age and sex
- M Named groups of persons
- N Health care concepts
- Q Biomedical disciplines, science and art

This structure is modelled on that of MeSH (see section 1). However, the distribution of terms differs from that of MeSH: Emtree - with its focus on drug names - has more drugs and chemicals in facet D (ca. 27,000 in 2009) than MeSH has total terms (ca. 25,000 in 2009).

Each facet represents a single term type, as described by the facet name. Within each facet, terms are organised into a hierarchical tree (up to 12 levels deep). Individual terms may appear more than once in the hierarchy, i.e. Emtree is polyhierarchical (explained below).

Emtree's hierarchical tree - the facet structure - is defined by over 7,500 explosion terms (preferred terms with at least one narrower term) which occupy over 12,000 locations ('nodes')<sup>1</sup>. The image below is a representation of the nodes in facets A-E. The facet with the most nodes is Facet C (Physical diseases, disorders and abnormalities), followed by facet D (Chemicals and drugs):



<sup>1</sup> In the printed Emtree, explosion terms have codes which identify their place in the tree structure. This enables users to find the hierarchical location of terms which they have identified in the alphabetical index.

Terms at different levels in the hierarchy have a broader-narrower term relationship with each other, as shown in the inset, which represents part of facet D (e.g. hydrolase inhibitor is a narrower term of the higher level term enzyme inhibitor).

Several additional aspects of the facet structure merit special attention:

- Polyhierarchical structure: terms may be listed in several different parts of the hierarchy. For example, leukemia is listed both in the malignant neoplastic disease (cancer) tree and in the leukocyte disorder tree. If such multiply-listed terms are explosion terms (i.e. head a hierarchy of narrower terms), Emtree displays the same hierarchy in all locations.
- Non-explosion terms: these are terms - 48,900 in all - which lack their own codes and which are associated as narrower terms with one or more explosion terms. In the inset, the explosion term renin inhibitor has 31 narrower non-explosion terms, including inhibitors such as ciprokiren and zankiren (not shown).
- Explosion searches: the tree structure allows users to do explosion searches to retrieve groups of hierarchically related terms in a single search. For example, an explosion search on proteinase inhibitor retrieves records indexed with either this term or any of its narrower terms, including aspartic proteinase inhibitor, renin inhibitor, as well as all narrower (non-explosion) terms of these terms (such as ciprokiren and zankiren).

If explosion searches are carried out using terms present in more than one place in the hierarchy (as in the case of leukemia, see above), then all narrower terms are retrieved.

## 4. Unique aspects of Emtree

### 4.1. Natural language terminology

In Emtree, terms are formatted whenever possible using natural language, i.e. as found in the article text. In contrast to MeSH, inverted forms are not used, e.g.

Emtree: *acute biphenotypic leukemia*  
MeSH: *leukemia, biphenotypic, acute*

Natural language terminology is an advantage for both indexers and searchers: indexers can assign as index terms the phrases they find in the text, and users can search intuitively, using phrases that they are familiar with (e.g. here: *acute biphenotypic leukemia*).

### 4.2. Scope Notes

Scope notes are not available for most terms. Since Emtree accepts as index terms the natural language phrases that indexers find in the text, users can be confident that Embase follows natural usage and does not impose restrictive definitions which might interfere with comprehensive retrieval.

Terms which might be ambiguous are not used. For example, there is no term *equilibrium* in Emtree. Instead, its multiple meanings are represented using the terms: *body equilibrium*, *acid base equilibrium* and *psychological equilibrium*.

The only Emtree terms which have scope notes are Check tags and Subheadings. In these cases, scope notes do supply important additional information defining term usage (see below).

### **4.3. Check tags**

Check tags are frequently used terms in categories such as *item type*, *age group* and *study type* (there are about 50 such terms in all). Most Embase records are indexed with several check tags, which can be used to limit searches when (otherwise) too many records would be retrieved.

Because of the importance of precise definitions in these cases, check tags have scope notes. For example, *adolescent* is defined (for humans) as between the ages of 13 and 17.

### **4.4. Subheadings (previously known as Link terms)**

There are two kinds of subheading: drug subheadings and disease subheadings. Subheadings are used to modify drug and disease terms respectively, and in their use closely resemble MeSH subheadings. Like check tags they also have scope notes.

Emtree has 64 drug subheadings (including 47 routes of drug administration) and 14 disease subheadings. Examples of their use are:

- drug subheading: *ranitidine (adverse drug reaction)*
- disease subheading: *arthritis (etiology)*

Typically, indexed drug and disease terms are each modified by several subheadings, which describe the context of use of these terms in Embase records. For example, subheadings differentiate records describing disease treatment (*disease linked to drug therapy*) from records in which disease symptoms arise as adverse reactions (*disease linked to side effect*).

### **4.5. Extensive synonyms**

As indicated above, Emtree is rich in synonyms. Synonyms include many spelling variants as well as alternative descriptions for preferred terms, thus providing users with as much help as possible in identifying the appropriate search term. In the case of drugs and chemicals, the following kinds of synonym are available:

- alternative generic names (e.g. USAN, BAN)
- alternative spelling (e.g. *acyclovir* is mapped to the preferred term *aciclovir*)
- chemical names
- trade names
- laboratory and research codes

On average, Emtree drug preferred terms each have more than 5 synonyms, which adds up to a total of more than 144,000 drug synonyms in Emtree.

### **4.6. MeSH terms**

Emtree includes all MeSH terms. Each year, all new MeSH terms from the preceding year are reviewed (many are already in Emtree) and appropriate assignments as preferred terms or (more typically) synonyms are identified.

This enables users who know MeSH to search Embase using the terminology that they are most familiar with.

#### **4.7. CAS registry numbers**

Emtree drugs are linked in a separate table to CAS registry numbers. This table is used to generate CAS registry numbers in Embase whenever the corresponding drug is indexed. CAS registry numbers help users do drug searches across different databases, where they may represent the only vocabulary of drug “names” common to all databases.

Some Emtree drugs are linked to multiple CAS registry numbers. The additional numbers belong to salt forms or isomers which are synonyms of the preferred term. Overall, more than 20,300 CAS registry numbers are associated with more than 16,700 preferred terms. The table is updated each year as more drugs are added to Emtree.

#### **4.8. Enzyme Commission (EC) numbers**

EC numbers are codes assigned to enzymes to facilitate retrieval across databases which may represent enzyme names in different ways. For example, the enzyme hydroxymethylglutaryl coenzyme A reductase has the EC code ec 1.1.1.88. These codes are included in Emtree as synonyms of the corresponding enzyme names.

### **5. Emtree maintenance**

New concepts - drugs, diseases, procedures and more – are continuously described in the biomedical literature, and potential new terms based upon these concepts are initially identified by Embase indexers. Each year, more than 200,000 candidate terms are identified in this way. Procedures designed to identify which candidate terms should be promoted into Emtree are described below. To keep up with these changes Emtree is updated every year, making Emtree one of the most current thesauri available, particularly for new drugs and diseases.

#### **5.1. New terms**

During the year, over 2000 drug candidate terms with the highest frequency of use are reviewed by Elsevier editorial staff, leading to the addition to Emtree of some 500 new drug preferred terms (with their CAS registry numbers) per year. Many synonyms, including trade names and codes identified in other sources such as the websites of pharma companies, are also added.

Similarly, new diseases, organisms and procedures, and to a lesser extent other terms such as geographical terms, are identified from the list of candidate terms, resulting each year in the addition of some 300 non-drug terms to Emtree. As with drug terms, the choice is made based upon importance and frequency of use.

In addition, each year new MeSH terms from the preceding year are added to Emtree as described above (see section 4.6).

#### **5.2. Changed terms**

In some cases, the new terms identified in the course of a year turn out to be new names for pre-existing terms that are already in Emtree. This is most typical for drug terminology. For example, in 1999 the following preferred term (with synonym *cs 866*) was introduced:

*4 (1 hydroxy 1 methylethyl) 2 propyl 1 [[2' (1h tetrazol 5 yl) 4 biphenylyl]methyl] 5 imidazolecarboxylic acid 5 methyl 2 oxo 1,3 dioxol 4 ylmethyl ester*

In 2001, this unwieldy chemical name was made synonym of the generic name *olmesartan*; the trade names *benevas*, *benicar*, *olmetec*, and *votum* were added later as new synonyms.

Note: Trade names and laboratory codes such as those listed here are usually synonyms in Emtree. When indexed in Embase, they are mapped to their preferred terms (usually the generic name), so that it is in general not possible to determine from the Emtree indexing which (if any) trade name was mentioned in the original article.

However, Embase also has a separate Trade name index that is not controlled against Emtree. By searching names in this index, it is possible to retrieve records which have been indexed with a specific Trade name.

### **5.3. *New structure***

The Emtree structure is reviewed annually, both in order to identify new term categories and to ensure that the existing categories are complete.

For example, in 2009 the term glucagon like peptide was made an explosion term with its naturally occurring and synthetic derivatives as narrower terms. This means that these compounds can now be searched as a group.

### **5.4. *Creation dates***

The year in which each preferred term was introduced in Emtree is documented as a Creation Year. This information is currently reproduced in the printed Emtree.

For online files, the Creation Date is the actual date on which the term was entered into the Emtree system.

## Appendix 1. Emtree Facts (2009)

Preferred Terms:	56,490 (drugs: 27,403; other: 29,087)
Synonyms:	230,662 (drugs: 144,846; other: 85,816)
Codes:	12,505 (drugs: 1,758; other: 10,747)
CAS Registry nrs:	20,392 (preferred terms with CAS nrs: 16,738)
MeSH terms:	24,626

### *Drug subfacets*

D1	general and inorganic chemicals
D2	organic compound
D3	pharmaceutical vehicles and additives
D4	natural products and their synthetic derivatives
D5	environmental, industrial and domestic chemicals
D6	hormones and agents acting on the endocrine system
D7	agents acting on the genital system
D8	urinary tract agent
D9	digestive tract agent
D10	respiratory tract agent
D11	agents acting on the auditory and vestibular systems
D12	agents acting on the eye
D13	dermatological agent
D14	analgesic, antiinflammatory, antirheumatic and antigout agents
D15	central nervous system agents
D16	agents interacting with transmitter, hormone or drug receptors
D17	agents acting on the peripheral nervous and neuromuscular systems
D18	cardiovascular agent
D19	hematologic agent
D20	antiinfective agent
D21	antiparasitic agent
D22	antineoplastic agent
D23	diagnostic agent
D24	biologic factors and agents acting on the immune system
D25	biomedical and dental materials
D26	agents used in emergency medicine
D27	drugs used in the treatment of addiction
D28	agents affecting water, molecule or ion transport
D29	agents affecting metabolism
D40	miscellaneous drugs and agents

Each subfacet (and/or its subdivisions) belongs itself to one of the following subcategories:

- structure (D1, D2)
- natural occurrence and analogs (D4)
- affected organ/system or therapeutic indication (D6 - D15, D17 - D22)
- mechanism of action (D16, D24, D28, D29)
- miscellaneous (D3, D5, D23, D25 - D27, D40)

These subcategories are used to classify drugs in Emtree applying the following hierarchy:

1. therapeutic indication; affected organ/system;
2. mechanism of action
3. natural occurrence

#### 4. chemical structure

Categories 1-3 above are always assigned if applicable; category 4 is assigned for most chemical names, and when the structure is a significant part of the mechanism of action or therapeutic category.

For example, ibuprofen has the following broader terms:

- antipyretic analgesic agent (therapeutic indication) (D14)
- nonsteroid antiinflammatory agent (therapeutic indication) (D14)
- prostaglandin synthase inhibitor (mechanism of action) (D29)
- arylpropionic acid derivative (chemical structure, characteristic for the '-profens') (D2)