MEDLINE®/PubMed® Baseline Repository (MBR) Reference Material

Date Last Updated: Thursday, June 01, 2017

1. Introduction

Researchers have requested the ability to have available MEDLINE citations in the state they were at a given moment in time without the MeSH vocabulary updates and other revisions that occur during the year. The MEDLINE/PubMed Baseline Repository was setup to provide this capability. We have stored the end of year baseline of the MEDLINE/PubMed database for each year starting in 2002.

We have also created tools to do some basic calculations like frequency counts of MeSH terms for each of the baselines.

2. Restrictions/Caveats

Users are responsible for compliance with the U.S. National Library of Medicine (NLM) Terms and Conditions which can be found at

https://www.nlm.nih.gov/databases/download/terms_and_conditions.html

The records included in the MEDLINE/PubMed Baseline databases **represent a static view of the data at the time each baseline database was created**.

3. Baselines

The baselines are normally generated towards the middle of November each year and contain all completed citations in MEDLINE as of that date. The baselines represent MEDLINE after the year-end processing has been completed. This means that the records have been revised with the upcoming year's new MeSH vocabulary terms. We currently have available all of the MEDLINE/PubMed Baselines from 2002. The naming of the baselines represents this year-end processing. For example, the 2002 MEDLINE/PubMed Baseline contains all completed citations from the mid-1960's until the date the baseline was created in late November 2001 with the year-end processing assigning appropriate 2002 MeSH vocabulary terms, thus it is a baseline for the 2002 year.

The baselines contain citations that are not MEDLINE as well. All of the baselines we have stored (2002 on) contain "Out-of-scope" citations which were renamed to "PubMed-not-MEDLINE" starting with the 2004 MEDLINE/PubMed Baseline. The PubMed-not-MEDLINE status refers to citations that reside in PubMed from journals included in MEDLINE and have undergone quality review but are not assigned MeSH headings because the cited item is not in scope for MEDLINE either by topic or by date of publication. Citations in the Out-of-scope or PubMed-not-MEDLINE status make up a very small percentage (0.51% or 75,271 records in the 2005 baseline) of the total number of citations contained in the baselines.

Starting with the 2005 MEDLINE/PubMed Baseline, OLDMEDLINE citations are also included in the baselines. The OLDMEDLINE citations make up approximately 11% of the total number of baseline citations. The OLDMEDLINE citations are from international biomedical journals covering the fields of medicine, preclinical sciences, and allied health sciences. The citations were originally printed in hardcopy indexes published prior to 1966. For additional information, please refer to the following URL:
http://www.nlm.nih.gov/databases/databases_oldmedline.html.

In the 2005 baseline, the subject indexing from the OLDMEDLINE citations were stored solely in the "Other Term" (or "OT") tagged fields and not the MeSH Terms (or MH) tagged fields. This means that searching the 2005 baseline from our MBR Query Tool via the MH field does not include any OLDMEDLINE citations. The only way

to include OLDMEDLINE records in the 2005 baseline is to do a timeframe query without specifying any field specific search criteria. Beginning with the 2006 baseline, Other Terms are starting to be mapped to current MeSH Terms so that searching via the MH field may retrieve some OLDMEDLINE records, but, not necessarily the complete set of possibilities.

MeSH Files: Starting in 2017, the NLM MeSH Section has made available the MeSH FTP download site: ftp://nlmpubs.nlm.nih.gov/online/mesh/ which includes separate directories for each release year of MeSH. In addition, they created the folder "MESH_FILES" with the latest release files that are updated every morning Monday - Friday. The yearly release folders span from 2011 to the latest full release which occurs in November of the preceding year (for example, 2016 MeSH was released in November of 2015). **A single directory is also included for earlier files from 1999-2010**.

For each release of MeSH, we create the **streeYYYY.bin.s** file which contains a mapping for each MeSH Heading to the Semantic Type(s) that are associated with that MeSH Heading.

DTD Files: We save a copy of the relevant DTD (Document Type Definition) files each year for working with the Baseline XML files.

Frequency Count Files: Basic frequency counts for the entire MEDLINE/PubMed Baseline sorted into alphabetical and numerical order for the following MEDLINE fields. For all fields but the NM field, we also provide a sort and count of their occurrences as starred (Index Medicus) items.

a. MH (MeSH Headings)
b. SH (MeSH Subheadings)
c. MH/SH combinations
d. NM (Chemicals)

Raw Data Files: Files containing the raw data for this Baseline year. There is a README file describing the various files available and their layouts.

Histogram/Summary Files: File showing the number of MH terms assigned to each of the various MeSH Tree top-level and top-level + 1 categories during the latest year to see how assignment of terms might vary from year to year. File showing the number of MH terms assigned to each of the UMLS Semantic Type Groupings categories during the latest year to see how assignment of terms might vary from year to year from a different perspective.

Frequency Counts:

Simple frequency counts were done of all MH and RN lines for all citations in each of the baseline collections. These frequency counts were then aggregated into the following categories: Chemical (RN terms), Main Heading (MH terms), SubHeading (qualifier terms), and Main Heading/SubHeading (MH/qualifier combination terms). The results are provided either in alphabetical or numerical sorted order for each category.

Chemical Files (RN terms):

-- Format of files:

Frequency count overall in MEDLINE
Registry Number for the Chemical
Chemical name

-- Chemical_freq_alpha -- Ordered by Chemical name
-- Chemical_freq_count -- Ordered by Frequency counts (high->low)

Main Heading (MH) Files (MH terms):

-- Format of files:

Frequency count overall in MEDLINE
Frequency count when starred (major) item
MeSH Heading

-- MH_freq_alpha -- Ordered by MeSH Headings
-- MH_freq_count -- Ordered by Overall Frequency counts (high->low)
-- MH_major_freq_count - Ordered by Starred Frequency counts (high->low)

SubHeading (SH) Files (qualifier terms):

-- Format of files:

Frequency count overall in MEDLINE
Frequency count when starred (major) item
SubHeading or Qualifier name

-- SH_freq_alpha -- Ordered by Subheadings/qualifier
-- SH_freq_count -- Ordered by Overall Frequency counts (high->low)
-- SH_major_freq_count - Ordered by Starred Frequency counts (high->low)

Main Heading/SubHeading (MH_SH) Combination Files (MH/qualifier combination terms):

-- Format of files:

Frequency count overall in MEDLINE
Frequency count when starred (major) item
MH/SubHeading name combination

-- MH_SH_freq_alpha -- Ordered by MH/SH
-- MH_SH_freq_count -- Ordered by Overall Frequency counts (high->low)
-- MH_SH_major_freq_count - Ordered by Starred Frequency counts (high->low)

Special Note: This counts each of the MH/SH combinations, applying the starred count only to combinations that include the starred term - for example:

If we have an entry like the following in a citation:

MH - Obesity/*complications/etiology

We include the following entries in this table:

Obesity/*complications
Obesity/etiology

4. MH Term Assignment by MeSH Treecodes Summary (hist)

The MH Term assignment summary was done using the information garnered from the frequency counts. We found the MeSH Tree code for each of the unique MH terms found for each baseline and then summarized the frequency counts for the top two MeSH Tree levels. The idea for this summary was to see if we could detect any patterns in the way MeSH terms are assigned from year to year by creating histograms comparing the various years either within a baseline or across baselines. The example below is a view from the 2004 MEDLINE Baseline summary file.

Top-Level Summary:

```
Anatomy [A]|357747
Organisms [B]|392381
Diseases [C]|537255
Chemicals and Drugs [D]|980930
Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]|877038
Psychiatry and Psychology [F]|185296
Biological Sciences [G]|1024627
Physical Sciences [H]|383770
Anthropology, Education, Sociology and Social Phenomena [I]|109299
Technology and Food and Beverages [J]|45889
Humanities [K]|27983
Information Science [L]|149576
Persons [M]|388012
Health Care [N]|529399
Geographic Locations [Z]|94882
```

Second-Level Summary:

```
Body Regions [A01]|14835
Musculoskeletal System [A02]|32999
Digestive System [A03]|20192
Respiratory System [A04]|8780
Urogenital System [A05]|18956
Endocrine System [A06]|8560
Cardiovascular System [A07]|30369
Nervous System [A08]|63992
Sense Organs [A09]|11155
Tissues [A10]|36849
…
Population Characteristics [N01]|59665
Health Care Facilities, Manpower, and Services [N02]|75251
Health Care Economics and Organizations [N03]|59176
Health Services Administration [N04]|106004
Health Care Quality, Access, and Evaluation [N05]|322311
Geographic Locations [Z01]|94882
```

5. MH Term Assignment by UMLS Semantic Type Group Summary (histST)

The MH Term assignment summary was done using the information garnered from the frequency counts. We found the UMLS Semantic Type(s) for each of the unique MH terms found for each baseline and then summarized the frequency counts for each of the UMLS Semantic Type Groups which are high-level categories combining multiple Semantic Types. The idea for this summary was another way to see if we could detect any patterns in the way MeSH terms are assigned from year to year by creating histograms comparing the various years either within a baseline or across baselines. The example below is a view from the 2004 MEDLINE Baseline summary file.

```
Activities & Behaviors|164011
Anatomy|364259
Chemicals & Drugs|1719653
Concepts & Ideas|642821
Devices|30862
Disorders|640501
Genes & Molecular Sequences|76533
Geographic Areas|95287
Living Beings|798945
Objects|65613
Occupations|71534
Organizations|29226
Phenomena|128390
Physiology|408765
Procedures|747397
```

6. Histogram/Summary Files

We have created a couple of histograms or summaries of how the MeSH Headings (MH) are assigned. The first perspective is from how the MHs are assigned based on their MeSH Vocabulary Treecodes. In the MeSH Vocabulary, each MH is assigned one or more treecodes, and we assign counts for all of the treecodes assigned. The second perspective is how the MHs are assigned based on their UMLS Semantic Types. Each UMLS concept (MHs are concepts in UMLS) is assigned one or more Semantic Types and we assign counts for all of the Semantic Types assigned. We then roll the individual counts for the Semantic Types up into the UMLS Semantic Groupings. We have also taken the time to graph the various histograms to provide a visual as well as data driven view of the data.

The Treecodes and Semantic Types help describe the MeSH Headings and place them appropriately within either MeSH or the UMLS. For more information on MeSH Treecodes, UMLS Semantic Types, and UMLS Semantic Groupings, please refer to the "References" section at the bottom of this page.

File	Description
hist	A count of MeSH Main Headings partitioned into the respective MeSH Treecodes for this year. We include counts for the top-level (e.g., "A", "B") and top-level plus one (e.g., "A01", "A02, "B01") MeSH Treecodes based on the MeSH Vocabulary specific to each baseline year.
hist.pdf	A graph generated from the respective hist file data.
histST	A count of UMLS Semantic Types partitioned into the respective high-level UMLS Semantic Groupings.
histST.pdf	A graph generated from the respective histST file data.
hist_Full	We use the same methodology we used in the above "hist" file, but, apply it across the entire MEDLINE and report counts for each of the years (1965 - 2003) in the MEDLINE Baseline. This file is provided for comparison purposes to see how assignments have changed over the years. **NOTE:** This file is created using the 2004 MeSH Vocabulary, so the numbers will differ from the 2002 and 2003 baseline counts.
hist_Full.pdf	A graph generated from the respective hist_Full file data.
histST_Full	We use the same methodology we used in the above "histST" file, but, apply it across the entire MEDLINE and report counts for each of the years (1965 - 2003) in the MEDLINE Baseline. This file is provided for comparison purposes to see how assignments have changed over the years. **NOTE:** This file is created using the 2004 MeSH Vocabulary, so the numbers will differ from the 2002 and 2003 baseline counts.
histST_Full.pdf	A graph generated from the respective histST_Full file data.
combined_hist.pdf	A graph generated from the "hist" methodology results for all of the MEDLINE Baselines currently in the Repository for comparison purposes.
combined_histST.pdf	A graph generated from the "histST" methodology results for all of the MEDLINE Baselines currently in the Repository for comparison purposes.

NOTES:

1. All counts are based solely on MeSH Headings.
2. We use the Date Completed date to determine inclusion in the counts for each baseline.
3. We are using the MeSH assignment of Semantic Type for the MeSH Headings wherever possible. Where a Semantic Type was not found in the MeSH dYYYY.bin file for a given MeSH Heading, we manually determined the Semantic Type(s) via the UMLS MRCON and MRSTY files.
4. The hist_Full and histST_Full files which report on the entire MEDLINE baseline were created using the latest MeSH Vocabulary, 2004 in this case. This means that some counts may be different in this overall file than what is seen in the 2002 and 2003 baseline count files. The 2004 MeSH was used to provide consistency with using the 2004 Baseline to create the two files.

7. MEDLINE/PubMed Baseline Repository Detailed Resources Information

This section of the paper details where, when, how, and what resources are used for creating and maintaining the MEDLINE/PubMed Baseline Repository.

MEDLINE/PubMed Baseline:

The MEDLINE/PubMed Baselines are generated each year by the U.S. National Library of Medicine (NLM). The baseline is typically generated towards the end of November each year and officially announced around the middle of December each year. For status and information pertaining to the MEDLINE/PubMed Baselines, you can look at The National Library of Medicine Data Distribution web site under the MEDLINE/PubMed Biomedical journal citations section: https://www.nlm.nih.gov/databases/download/data_distrib_main.html.

UMLS Semantic Types Grouping (SemGroups.txt) file:

The SemGroups.txt file is the latest addition to the Repository and it's unclear whether this file is updated each year, as the Semantic Types change, or is static. This file has grouped the UMLS Semantic Types into 15 (currently) high-level categories. We are using this file to see if we can detect patterns in how the MeSH Headings are assigned in MEDLINE. The papers: "***Aggregating UMLS semantic types for reducing conceptual complexity*****.** McCray AT, Burgun A, Bodenreider O; Medinfo. 2001;10(Pt 1):216-20." and "***Exploring semantic groups through visual approaches*****.**, Bodenreider O, McCray AT; Journal of Biomedical Informatics. 2003; 36(6):414-432." provide much greater detail on the grouping of the Semantic Types. Both papers can be found at the Lister Hill National Center for Biomedical Communications web site (http://lhncbc.nlm.nih.gov).

To retrieve the SemGroups.txt file, we go to the UMLSKS (Unified Medical Language System Knowledge Source Server) at http://umlsks.nlm.nih.gov, login, and then go to the "Semantic Groups" link under the "UMLSKS Resources" section of the welcome page.