Fingerprinting Biomedical Terminologies – Automatic Classification and Visualization of Biomedical Vocabularies through UMLS Semantic Group Profiles

Bastien Rance[a], Thai Le[b], Olivier Bodenreider[c]

[a] *AP-HP, University Hospital Georges Pompidou; INSERM, UMR_S 1138, Centre de Recherche des Cordeliers, Paris, France*
[b] *Biomedical and Health Informatics, University of Washington, Seattle, WA, USA*
[c] *National Library of Medicine, National Institutes of Health, Bethesda, MD, USA*

Abstract

***Objectives**: To explore automatic methods for the classification of biomedical vocabularies based on their content. **Methods**: We create semantic group profiles for each source vocabulary in the UMLS and compare the vectors using a Euclidian distance. We explore several techniques for visualizing individual semantic group profiles and the entire distance matrix, including donut pie charts, heatmaps, dendrograms and networks. **Results**: We provide donut pie charts for individual source vocabularies, as well as a heatmap, dendrogram and network for a subset of 78 vocabularies from the UMLS. **Conclusions**: Our approach to fingerprinting biomedical terminologies is completely automated and can easily be applied to all source vocabularies in the UMLS, including upcoming versions of the UMLS. It supports the exploration, selection and comparison of the biomedical terminologies integrated into the UMLS. The visualizations are available at (http://mor.nlm.nih.gov/pubs/supp/2015-medinfo-br/index.html)*

Keywords:

UMLS; semantic groups; terminology fingerprinting; content-based classification

Introduction

The Unified Medical Language System® (UMLS) is a terminology integration system [1]. It provides broad coverage of the biomedical domain, from disorders to procedures to drugs to anatomical structures. While some source vocabularies focus on a subdomain of biomedicine (e.g., RxNorm for drugs), others, such as SNOMED CT and the NCI Thesaurus, provide coverage across biomedicine. However, selecting a biomedical terminology remains challenging for users, because there is no description of content coverage, i.e., no description of which subdomains are covered by a given terminology.

The UMLS used to provide a classification of source vocabularies based on usage. This classification, performed manually, leveraged the Medical Subject Headings (MeSH). This classification was heterogeneous, as it mixed usage and content categories. For example, categories such as “Nursing” and “Complementary Therapies” reflect usage, whereas the categories “Disease” and “Procedures” are based on content. Moreover, classification by usage does not necessarily align with classification by content. For example, the International Classification for Nursing Practice (ICNP®) and Nursing Interventions Classification (NIC) are both “Nursing” terminologies, although NIC predominantly contains therapeutic procedures, while ICNP also contains content about diagnoses and outcomes. In addition, source vocabularies may need to be classified into more than one category. Another limitation of this classification is that only the most frequently updated sources in the Metathesaurus were considered, because manual classification is labor-intensive. Overall, while useful to new users, this classification was imperfect and difficult to maintain for new versions of the UMLS in which new terminologies may have been introduced or modified significantly.

The objective of this work is to explore automatic methods for the classification of biomedical vocabularies based on their content. More specifically, we create a “fingerprint” (i.e., semantic profile) for each terminology in the UMLS by leveraging the categorization of UMLS concepts into semantic groups. These semantic group profiles form the basis for classifying and comparing biomedical vocabularies based on their content, and are expected to help users explore, select and compare terminologies (e.g., for text annotation purposes). Our approach is fully automatic, does not require any additional knowledge about the vocabularies, and can be easily deployed. We also explore several visualization techniques to render this classification. The semantic fingerprints we provide for biomedical terminologies could complement, if not replace, the classification of UMLS source vocabularies provided earlier.

Background

The Unified Medical Language System. The Unified Medical Language System® (UMLS) is assembled by integrating 179 source vocabularies. The UMLS Metathesaurus (version 2014AB) currently contains about 3.1 million concepts, i.e., clusters of synonymous terms coming from various source vocabularies. Each Metathesaurus concept is assigned at least one semantic type from the UMLS Semantic Network, a small network of 133 semantic types organized into a tree structure. The semantic types are partitioned into fifteen semantic groups (McCray et al. 2001), which represent broad subdomains of biomedicine, such as *Anatomy*, *Chemicals & Drugs*, and *Disorders*. Every semantic type is categorized into only one semantic group. The fifteen semantic groups are listed in Table 1, along with and the number of Metathesaurus concepts in each group. In practice, the semantic groups provide a coarse categorization of the Metathesaurus concepts based on the principles of semantic validity, parsimony, completeness, exclusivity, naturalness, and utility. The semantic groups have been used in several applications, including visualization of highly conceptual spaces [2], discovery of inconsistencies in the categorization of UMLS concepts [3], word-sense disambiguation [4], and quality assurance of value sets [5].

Visualization and cognition. We present different graphical representations of vocabularies based on their semantic content. There exists a broad body of literature describing the impact of visual displays on not only the speed of decision-making, but also its accuracy [6-9]. Cognitive theories provide context on the processes involved in visualizing information. This can range from the theories of Cleveland and McGill, who propose a set of elementary visual tasks for interpreting displays, to Pinker's models of cognitive processing from raw visual information to encoded visual descriptions [10, 11]. We also leverage some of the visualization techniques used for genomic datasets [12]. In the context of fingerprinting biomedical terminologies, visualization is an important component of presenting the information in a succinct manner to facilitate use by a broad range of stakeholders within the biomedical community.

Table 1– Distribution of Metathesaurus concepts by semantic groups

Semantic group name	Abbreviation	# concepts
Activities & Behaviors	ACTI	4,385
Anatomy	ANAT	122,298
Chemicals & Drugs	CHEM	813,426
Concepts & Ideas	CONC	48,711
Devices	DEVI	45,883
Disorders	DISO	544,829
Genes & Molecular Sequences	GENE	67,760
Geographic Areas	GEOG	4,426
Living Beings	LIVB	948,012
Objects	OBJC	16,175
Occupations	OCCU	1,506
Organizations	ORGA	2,220
Phenomena	PHEN	12,778
Physiology	PHYS	140,146
Procedures	PROC	374,195

Methods

Our method for classifying biomedical vocabularies based on their content can be summarized as follows. For each source vocabulary in the UMLS, we first create a vector reflecting the distribution of its concepts among semantic groups (i.e., a semantic group profile). We then compare these semantic group profiles using a Euclidian distance. Finally, we apply several visualization techniques to the semantic group profiles.

Creating semantic group profiles

For each UMLS source vocabulary, we compute the frequency distribution of its concepts among the 15 semantic groups, which we record in a 15-dimensional vector. This is what we call the semantic group profile (or semantic fingerprint) of a source vocabulary. For example, SNOMED CT spans a variety of semantic groups, including *Disorders* (31%), *Chemicals & Drugs* (23), *Procedures* (11%), *Anatomy* (7%) and *Devices* (3%). In contrast, 99% of the concepts from the Foundational Model of Anatomy (FMA) belong to the semantic group *Anatomy*. Its semantic profile is sparse, with few semantic groups other than *Anatomy* having a value other than 0. The set of vectors computed for each terminology forms a matrix of terminologies by semantic groups.

Comparing semantic group profiles

In order to compare two semantic group profiles, we use a Euclidian distance metric, that is, the straight line distance between two vectors, i.e., between two semantic group profiles. (We also tested other similarity metrics including cosine, Jaccard, and Dice. However, the Euclidian distance provided a range of values more suitable for defining groups of source vocabularies using hierarchical clustering.) We generate a distance matrix by calculating the Euclidian distance between terminologies pairwise.

We then use an agglomerative method of hierarchical clustering to group together similar semantic group profiles. The agglomerative hierarchical clustering algorithm starts with a distance matrix and identifies the pair of source vocabularies that are the most similar. This forms the first cluster. The distance matrix is then recalculated, with complete linkage defining the distance between clusters as the largest distance between any two of its elements. The elements of the matrix are compared to find the next closest pair between sources or clusters. This is repeated until a single agglomerative cluster of all source vocabularies is formed.

Visualizing semantic group profiles

We propose three different visualizations for the semantic group profiles depending on what we want to emphasize. Namely, we visualize single semantic fingerprints (i.e., single terminologies) with "donut" pie charts, sets of semantic fingerprints (i.e., multiple terminologies) with heatmaps, and associations between terminologies and semantic groups with network representations.

Visualizing single semantic group profiles

We use ***"donut" pie charts*** for visualizing single semantic group profiles. In this visualization, the source is represented as a ring. The source ring contains arcs corresponding to each semantic group. The size of the arc is proportional to the size of the corresponding semantic groups in the source. In addition to displaying the profile of a give terminology, this representation also makes it easy to compare different profiles.

Visualizing sets of semantic group profiles

We provide a ***heatmap*** representation of the data found in the distance matrix. The source vocabularies are listed on the x-axis and the semantic groups are listed on the y-axis. Density on the heatmap corresponds to the percentage of all concepts within a source vocabulary that is found in a given semantic group. Density is color coded, with red for high percentages of a semantic group in terminology and yellow for low percentages. As a result, scanning a vertical slice (column) of the heatmap provides a visual representation of the semantic group profile for a source vocabulary. Conversely, by scanning a horizontal slice (row) of the heatmap, a user can easily identify those source vocabularies with a large proportion of concepts for this semantic group.

Additionally, we generate a ***dendrogram*** to visualize the hierarchical clustering of the source vocabularies. Short branches on the tree represent terminologies with similar semantic group profiles, while long branches represent more dissimilar source vocabularies. The dendrogram can be cut to obtain a given number of clusters. Clustering also helps contrast groups of terminologies with similar semantic group profiles within groups and different profiles across groups. An arbitrary number of clusters can be produced, depending on the threshold of similarity among clusters used.

Visualizing associations among terminologies

In order to visualize associations among terminologies through semantic groups, we apply a ***bipartite network visualization*** to the semantic group profiles for visual display of content across multiple source vocabularies.

- Nodes represent semantic groups on the one hand and source vocabularies on the other
- An edge from source S to semantic group G is drawn if the source vocabulary contains at least some percentage of concepts in G.

The network representation makes it easy to identify which source vocabularies share a high concentration of a particular semantic group, as these vocabularies all have edges to this semantic group. Different networks can be obtained by selecting different thresholds for the minimum proportion of concepts from semantic groups.

Implementation.

All statistical analyses and heatmap visualization were performed using the R statistical software. Network display and “donut” pie chart leverage the JavaScript library "D3 for Data-Driven Documents".

Results

Visualizing single semantic group profiles – “donut” pie charts

Figure 1 shows the semantic group profiles of four UMLS source vocabularies. As mentioned earlier, the Foundational Model of Anatomy (FMA) contains almost exclusively anatomical concepts, displayed in light green. Similarly, the Online Mendelian Inheritance in Man (OMIM) vocabulary essentially contains gene (green) and disease (red) concepts. In contrast, SNOMED CT, a general clinical terminology, contains concepts from almost all semantic groups, with a large proportion of disease concepts. Finally, while the National Drug File-Reference Terminology (NDFRT) is a drug terminology, it also contains not only a majority of drug concepts (dark green), but also large numbers of concepts from other semantic groups, including *Disorders* (red) and *Physiology* (orange), because NDF-RT drugs are described in terms of physiologic effect and mechanism of action (*Physiology*), as well as therapeutic intent (*Disorders*).

Visualizing sets of semantic group profiles – heatmaps and dendrograms

Figure 2 shows the heatmap and dendrogram resulting from the hierarchical clustering of 78 source vocabularies. (Although the distance matrix was computed for all source vocabularies, the display is limited to these 78 vocabularies for readability. In practice, we filtered out non-English vocabularies. While translations of vocabularies contain new labels for concepts, their semantic content is identical to that of their English source. We also ignored vocabularies with fewer than 1,000 concepts since their small size limits their overall significance.)

Columns from the heatmap represent the semantic group profiles of individual source vocabularies. For example, the Foundational Model of Anatomy is represented by a single red spot for the semantic group *Anatomy*, while SNOMED CT spans multiple semantic groups in the column. Conversely, the rows of the heatmap reflect the density in concepts from a given semantic group. The large red bar in the lower right corner corresponds to a high density of concepts from the *Disorders* semantic group in disease terminologies.

The clustering algorithm was (arbitrarily) required to produce 6 clusters. Each cluster is rooted by the top subdivisions of the dendrogram (and highlighted by boxes with solid lines on the figure). Clusters range in size from 1 source vocabulary (HGNC) for the leftmost cluster, to 26 source vocabularies for the rightmost cluster. Some clusters are homogenous. For example, cluster 1 contains one gene terminology, cluster 2 contains 6 procedure terminologies and cluster 4 contains two terminologies primarily containing organisms. In contrast, the remaining clusters are heterogeneous and subgroups can easily be identified within them. For example, the large cluster 3 groups drug terminologies such as RxNorm, device terminologies, such as the Current Procedural Terminology (CPT), and general terminologies, such as SNOMED CT. Similarly, cluster 5 groups organism terminologies, such as the NCBI Taxonomy, anatomical terminologies, such as the Foundational Model of Anatomy (FMA), administrative terminologies, such as the HL7 value sets (HL7V2.5), and terminologies with focus on physiological concepts, such as LOINC and the International Classification of Functioning (ICF). Finally, cluster 6 clearly groups disease terminologies, some of which contain only disease concepts (e.g., ICD 10-CM), while others also contain concepts from other groups (e.g., genes and diseases in OMIM).

Visualizing associations among terminologies – networks

The bipartite network we created for visualizing associations among terminologies contains two types of nodes. The source vocabularies are represented in green, while the semantic groups are in yellow. Edges are drawn between a source vocabulary and a semantic group if the vocabulary contains at least 5% of concepts from this semantic group. (This arbitrary threshold can be modified to reflect stronger associations.) In Figure 3, source vocabularies that contain at east 5% of concepts from the semantic group *Disorders* are highlighted. Similarly, as shown in the inset from Figure 3, it is also possible to highlight all semantic groups for a given source vocabulary (i.e., all the semantic groups, whose concepts constitute at least 5% of the source vocabulary).

Discussion

Use cases and applications

The semantic group profiles provide a method for assessing the similarity among source vocabularies in the UMLS. This general technique can be applied to terminology exploration, terminology selection and terminology comparison.

Exploring terminologies

Novice users of the UMLS sometimes have difficulties grasping the differences among the many source vocabularies in the Metathesaurus. While the UMLS Terminology Services browser allows users to find the details about individual Metathesaurus concepts and their relations, it does not provide an overview of sets of concepts in source vocabularies. Our donut pie charts, heatmap and network visualizations provide an overview of the content of the source vocabularies. More specifically, they provide a coarse description of the semantics of these terminologies, making it possible to quickly identify the major semantic areas in a given vocabulary.

Selecting terminologies

One common use case is to select the best terminology for a given application. For example, if an application requires disease concepts, our visualizations make it easier for a user to identify candidate terminologies, i.e., terminologies containing a large proportion of concepts from the semantic group *Disorders*. In practice, a user will look for a large red arc on the donut pie charts, or might scan the *DISO* row on the heatmap, looking for red spots. Alternatively, our user could also select the *DISO* node on the network visualization and explore all source vocabularies linked to it, having set an appropriate threshold for the minimal proportion of concepts from this semantic group required for edges to be drawn.

Comparing terminologies

The heatmap is also the visualization of choice for analyzing sets of source vocabularies, especially after the hierarchical clustering has grouped together those terminologies that have similar semantic group profiles. The clusters displayed in Figure 2 and presented in the Results section are relatively easy to interpret, with minimal prior knowledge of the terminologies themselves. Similarity clusters can also be quantified, since the basis for clustering is the Euclidian distance computed among the semantic group profiles for individual source vocabularies.

Content-based vs. usage-based classification

Our work was motivated in part by the limitations of the usage-based classification the UMLS documentation used to provide. We compared the two classification approaches for the 55 source vocabularies for which it was available. The general trend is that there is limited overlap between the two classifications. Categories from the usage-based classification are generally associated with several semantic groups, and a given semantic group is generally associated with multiple categories from the usage-based classification, with no obvious patterns in these associations. One exception is the usage-based category "Adverse drug reaction reporting" that contained only one source vocabulary (MedDRA) and majoritarily contains concepts from the semantic group disorders. In fact, the two classifications provide different views on the source vocabularies and are complemantary. For example, it would be impossible to identify consumer health vocabularies or nursing vocabularies simply from the semantic group profiles. However, as mentioned earlier, unlike the manual usage-based classification, our semantic group profiles can be applied automatically to any new version of the UMLS. Finally, another advantage of our method is that, because it is a vector-based representation of the source vocabularies, it lends itself nicely to visual representation.

Limitations

Many concepts have more than one semantic type; however, these multiple semantic types are generally categorized into the same semantic group. Therefore most concepts are categorized by only one semantic group. In fact, only about 1,000 concepts have multiple semantic groups. As a result, the fifteen semantic groups form partition for over 99.9% of all UMLS concepts, and are thus virtually disjoint. For the purpose of computing the distribution of the concepts from a source vocabulary into semantic groups, the concepts that have multiple semantic groups should logically not be counted more than once. In practice, these concepts are so few in the UMLS that double-counting them has no significant effect on the frequency distributions.

The assignment of a semantic type to a UMLS concept is sometimes subjective and can be arguable. Many concepts are categorized with multiple semantic types. In contrast, all UMLS concepts are categorized in 15 disjoint semantic groups. Because the semantic groups are broader, the assignment of concept to a group is less likely to be arguable. However, some groups can be viewed as too general for this application. For example, the semantic group "Chemicals" contains both drugs and other chemicals. A user could be interested in retrieving drug vocabularies, rather all chemical vocabularies. As suggested in [13], the grouping of semantic types into semantic groups could be modified to fit the requirements of a particular application.

Future work

In this study we have used our fingerprinting methodology on UMLS source vocabularies. Leveraging concept mappings among terminologies, our approach could be used to automatically classifiy the content of non-UMLS terminologies in repositories such as the NCBO Bioportal. Our semantic group profiles could also be used to help characterize resources annotated to UMLS concepts, e.g., biomedical articles or clinical text annotated by MetaMap.

Conclusion

The growth of the UMLS makes it difficult for users to select appropriate source vocabularies for a given purpose. In this article, we present a new method to classify biomedical terminologies based on their content. We leverage the high level semantic categorization of concepts in semantic groups to create a profile for each source vocabulary. Our approach is completely automated and can easily be applied to all source vocabularies in the UMLS, including upcoming versions of the UMLS.

To assist the user in the exploration of available source vocabularies, we propose several visualizations reflecting the individual content of source vocabularies (donut pie charts, heatmaps), as well as the relations among source vocabularies (dendrogram, network). We are currently collaborating with the UMLS team to add the graphical representations to the UMLS documentation, as a complement to the classification they already provide.

References

[1] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;32(Database issue):D267-70

[2] Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. *J Biomed Inform* 2003;36(6):414-32

[3] Mougin F, Bodenreider O, Burgun A. Analyzing polysemous concepts from a clinical perspective: application to auditing concept categorization in the UMLS. *J Biomed Inform* 2009;42(3):440-51.

[4] Jimeno-Yepes A, McInnes BT, Aronson AR. Collocation analysis for UMLS knowledge-based word sense disambiguation. *BMC Bioinformatics* 2011;12 Suppl 3:S4.

[5] Jiang G, Solbrig HR, Chute CG. Quality evaluation of value sets from cancer study common data elements using the UMLS semantic groups. *J Am Med Inform Assoc* 2012;19(e1):e129-36.

[6] Elting LS, Martin CG, Cantor SB, Rubenstein EB. Influence of data display formats on physician investigators' decisions to stop clinical trials: prospective trial with repeated measures. *BMJ* 1999;318(7197):1527-31

[7] Feldman-Stewart D, Brundage MD, Zotov V. Further insight into the perception of quantitative information: judgments of gist in treatment decisions. *Med Decis Making* 2007;27(1):34-43

[8] Hoeke JO, Bonke B, van Strik R, Gelsema ES. Evaluation of techniques for the presentation of laboratory data: support of pattern recognition. *Methods Inf Med* 2000;39(1):88-92

[9] Morrow DG, Hier CM, Menard WE, Leirer VO. Icons improve older and younger adults' comprehension of medication information. *J Gerontol B Psychol Sci Soc Sci* 1998;53(4):P240-54

[10] Cleveland W, McGill R. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association* 1984;79(387):531-554

[11] Pinker S. A theory of graph comprehension. In: Friedle R, editor. *Artificial intelligence and the future of testing*. Hillsdale, NJ: Erlbaum; 1990

[12] Schroeder MP, Gonzalez-Perez A, Lopez-Bigas N. Visualizing multidimensional cancer genomics data.

Genome Med 2013;5(1):9 http://www.ncbi.nlm.nih.gov/pubmed/23363777.

[13] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. *Stud Health Technol Inform* 2001;84(Pt 1):216-20

Acknowledgments

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine. BR is supported in part by the the Cancer Research and Personalized Medicine - CARPEM project (Cancer Research Site - SIRIC). The authors thank Steve Emrick for useful discussions and feedback on the semantic fingerprints of UMLS source vocabularies.

Address for correspondence

Corresponding author: olivier@nlm.nih.gov

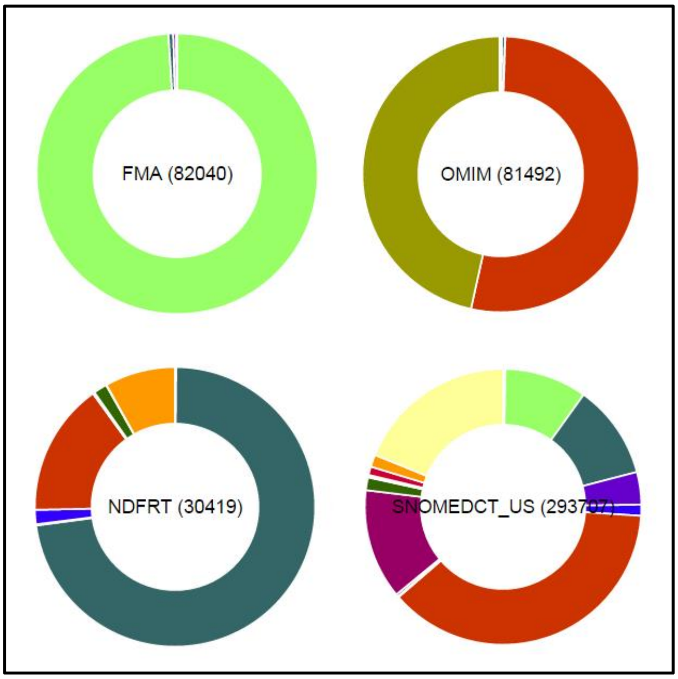

Figure 1– "Donut" pie charts for 4 UMLS source vocabularies. Color code: ANAT (light green), GENE (green), DISO (red), CHEM (dark green), PHYS (orange), LIVB (magenta), PROC (light yellow)

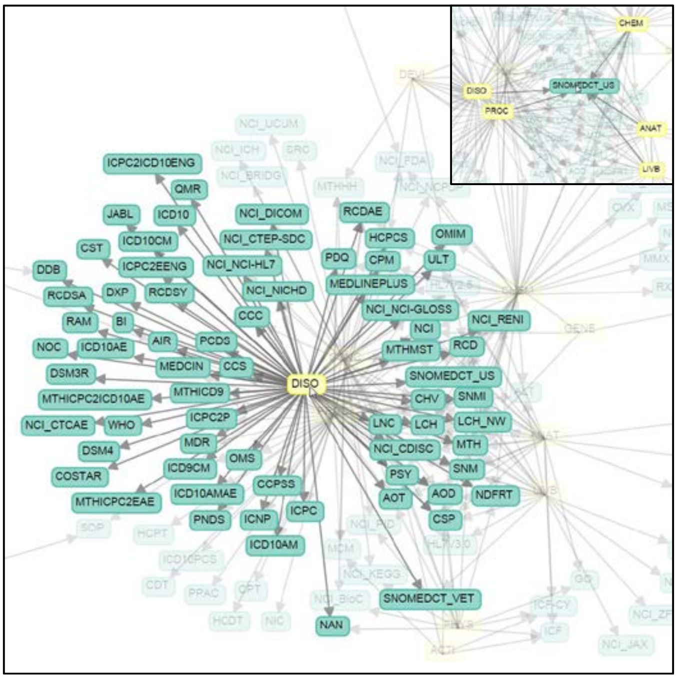

Figure 3– Network visualization of UMLS source vocabularies (green) linked to the semantic group Disorders (yellow) [Inset: Network visualization of SNOMED CT and its associations with several semantic groups]

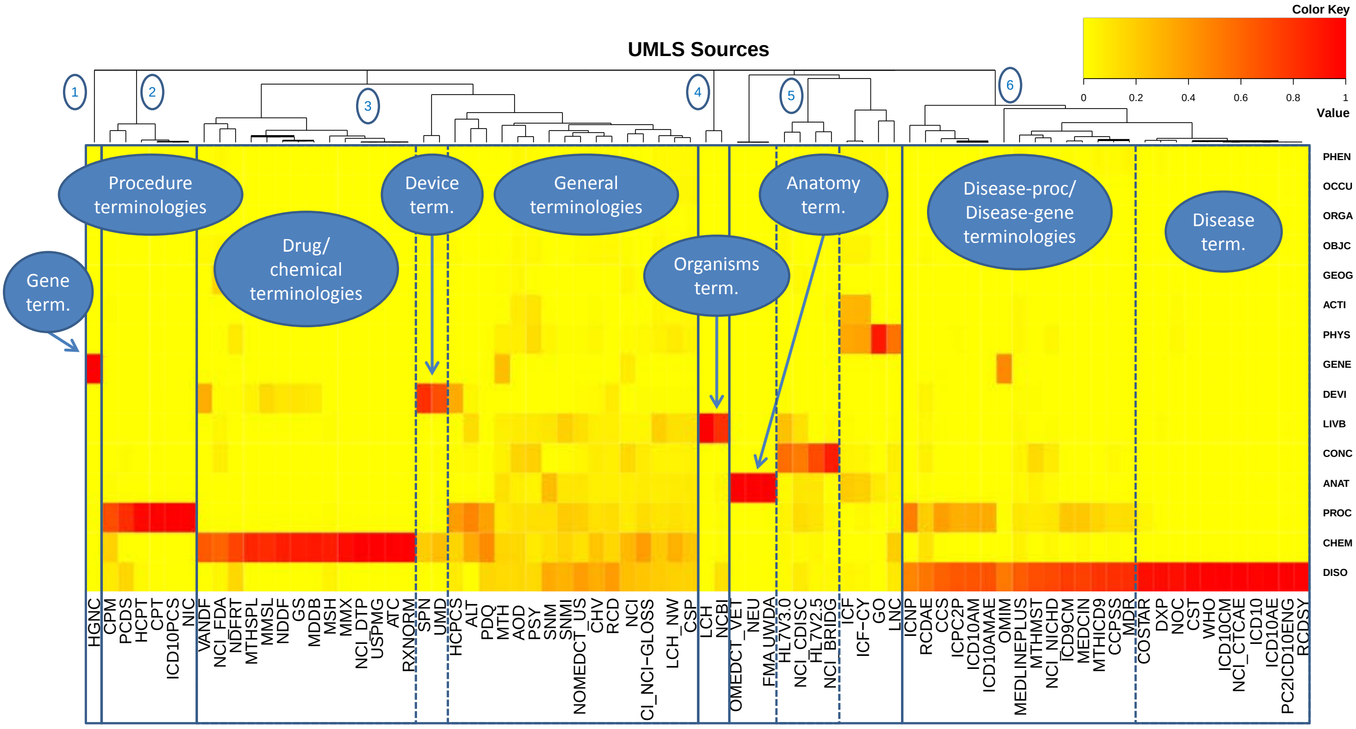

Figure 2– Heatmap of the UMLS terminologies and semantic groups. Bright yellow corresponds to the absence of a semantic group in a terminology. In contrast, bright red denotes a high percentage of concepts from the corresponding semantic group in the terminology.