Strength in Numbers: Exploring Redundancy in Hierarchical Relations across Biomedical Terminologies

Olivier Bodenreider, M.D., Ph.D.

U.S. National Library of Medicine, Bethesda, Maryland
National Institutes of Health, Department of Health & Human Services

olivier@nlm.nih.gov

Objectives*: To investigate three aspects of the redundancy of hierarchical relations across biomedical terminologies: 1) What proportion of the relations is redundant?, 2) Which terminologies tend to overlap with other terminologies?, and 3) Is there a link between redundancy and semantic consistency?.* ***Methods****: Hierarchical relations are counted in the various families of terminologies integrated into the UMLS and an index of redundancy is computed for each relation. Similarity among sources is computed using the classical cosine method. Semantic consistency is evaluated by reference to the UMLS Semantic Network.* ***Results****: Overall, 29% of the 1,128,261 relations examined exhibit redundancy. Most similar sources include consecutive versions of terminologies. The link between redundancy and semantic consistency is weak.* ***Discussion****: Applications of these findings are discussed, including selecting sources, selecting useful relations, and auditing the categorization of UMLS concepts.*

INTRODUCTION

Redundancy in biomedical terminologies has been considered essentially from the perspective of the concepts [1, 2]. Providing multiple names for a concept (i.e., synonymy) is generally considered a valuable feature [3], while multiple ways of representing a concept (e.g., through compositionality) should be avoided (unless the system allows equivalent expression to be recognized as such at the application level) [4]. At the same time, most authors favor multiple levels of granularity for concepts and multiple categorization of the concepts (resulting in multiple hierarchies) [1, 2].

In practice, these two features contribute to creating multiple paths between two concepts. For example, one path from *Pulmonary tuberculosis* to *Disease* may include *Lung disease*, while another includes *Infectious disease* (multiple inheritance). Moreover, the concept *Mycobacterium infection* may intervene between *Pulmonary tuberculosis* and *Infectious disease* in a terminology providing a higher level of granularity. The existence of multiple paths between two concepts is of course compounded when several terminologies are merged to form a broad terminological system such as the Unified Medical Language System® (UMLS®) Metathesaurus®.

From the perspective of relations, the existence of multiple paths between two concepts can be regarded as a different form of redundancy. The relation (C_1, parent of, C_2) may be considered redundant if it is found in several terminologies or if it can be inferred by combining several other relations, e.g., (C_1, parent of, C_3) and (C_3, parent of, C_2).

The objective of this experiment is to explore the redundancy of hierarchical relations in an inherently redundant terminological system: the UMLS Metathesaurus. More precisely, we want to address the following three aspects of redundancy in hierarchical relations in the Metathesaurus:

1) What proportion of the relations is redundant?
2) Which terminologies tend to overlap with other terminologies? (in terms of relations)
3) Is there a link between redundancy and semantic consistency?

We show that knowledge about redundancy in hierarchical relations may help customize a terminological system for various kinds of applications.

MATERIALS

The terminological system evaluated in this study is the Unified Medical Language System (UMLS), developed and maintained by the National Library of Medicine. The UMLS Metathesaurus[1] (13th edition, 2002AA) contains over 775,000 concepts from some sixty families of biomedical terminologies and over ten million relations (i.e., pairs of related concepts). As for the concepts, each relation may come from one or more sources. Nearly 1.2 million of these relations correspond to hierarchical relations contributed by the constituent terminologies or added by the Metathesaurus editors – namely (C_1, parent of, C_2) and (C_1, broader than, C_2) in Metathesaurus parlance). In order to benefit from the properties of directed acyclic graphs, we used a slightly modified version of the Metathesaurus from which the circular hierarchical relations have been removed [5].

[1] umlsinfo.nlm.nih.gov

1,155,673 hierarchical relations remained after this process was applied. In the Metathesaurus, each concept is categorized by means of semantic types from the Semantic Network. As mentioned in several studies [e.g., 6], this feature makes it possible to check the semantic validity of a hierarchical relationship between two concepts by comparing it to the relationships represented between the semantic types of the two concepts in the Semantic Network.

METHODS

Prior to investigating the three questions asked in the introduction, we must present what criteria we used for defining families of terminologies, redundancy, and semantic consistency.

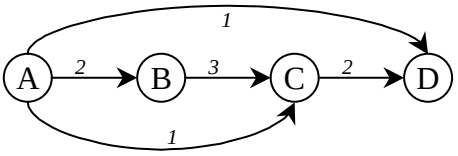
Definitions

Families of terminologies. In the UMLS, the constituent vocabularies are grouped by family[2]. For example, all translations of MeSH are part of the "MeSH family", identified by 'MSH'. Except for minor differences, we used the same grouping in this study and we refer the reader to the UMLS documentation for the full name of the source vocabularies. Forty-three families of terminologies contribute relations to the Metathesaurus.

Redundancy. The intuitive notion of redundancy for a relation is that of a relation shared by several sources. The redundancy for a given relation would thus be proportional to the number of sources providing this relation. However, this definition does not account for differences in granularity across terminologies or multiple categorization. Indeed, the pairs of hierarchically related concepts (C_1, C_2) and (C_2, C_3) can be seen as redundant with the pair (C_1, C_3). Moreover, the pairs (C_1, C_4) and (C_4, C_3) would also be redundant with (C_1, C_3). Thus, redundancy for (C_i, C_j) can rather be defined in terms of number of paths between C_i and C_j.

The index of redundancy for a given pair (C_i, C_j) is defined as the sum of the indexes of redundancy for each path between C_i and C_j. The index of redundancy for a given path is the minimum number of sources for each pair of concepts along the path ("weakest link" approach). As illustrated in Figure 1, the index of redundancy for (A, D) may be significantly higher than the number of sources for the direct relation between the two concepts.

In this experiment, we do not distinguish between the several types of hierarchical relationships in the Metathesaurus and a hierarchical relation is considered either present or absent in a source, regardless of its type in the source (parent, broader than, or both).



$IR_{AD} = IRP_{AD} + IRP_{ACD} + IRP_{ABCD} = 4$

with:

$IRP_{AD} = NS_{AD} = 1$

$IRP_{ACD} = \min(NS_{AC}, NS_{CD}) = \min(1, 2) = 1$

$IRP_{ABCD} = \min(NS_{AB}, NS_{BC}, NS_{CD}) = \min(2, 3, 2) = 2$

Figure 1 – Index of redundancy for (A, D).

Semantic consistency. Two concepts in hierarchical relationship, C_a (ancestor) and C_d (descendant), are considered semantically consistent if at least one of the following four situations occurs between their semantic types ST_a and ST_d:

1) ST_a and ST_d are identical;
2) ST_a and ST_d in taxonomic relation (i.e., ST_d isa ST_a);
3) ST_a and ST_d in partitive relation (i.e., ST_d part_of ST_a); or
4) ST_a and ST_d belong to the same semantic group[3].

When concepts have several semantic types, as they often do, the requirement for semantic consistency between C_a and C_d is that at least one semantic type of C_a be in one of the four situations above with at least one semantic type of C_d.

Q. 1. Counting redundant relations

For each pair of hierarchically concepts (C_1, C_2) in the Metathesaurus, we first computed the number of families of terminologies for the direct relation between C_i and C_j. Exploring all possible paths between C_i and C_j, we then determined the index of redundancy for the relation of C_i to C_j as defined earlier.

Q. 2. Measuring similarity among terminologies

We applied the vector space model used in information retrieval for measuring the similarity among documents to measuring the similarity among families of terminologies. For each family, we built a long vector of which each element corresponds to a hierarchical relation, valued 1 if this relation is present in the family and 0 otherwise. We normalized the vectors to compensate for unequal numbers of relations across terminologies. The cosine and Jaccard methods [8] were used to measure the similarity among vectors. In both cases, values close to 1 denote similarity.

[2] Family information is located in the column *SF* of the table *MRSAB*, recently added to the UMLS distribution

[3] cluster of semantically (but not necessarily structurally) related semantic types [7]

Q. 3. Comparing redundancy and semantic consistency

Having computed the index of redundancy for each pair of hierarchically related concepts (see Q. 1 above), we then determined whether the hierarchical relation between the two concepts in each pair was semantically consistent with corresponding relations in the Semantic Network, using the method defined earlier.

We purposely excluded from the comparison pairs whose ancestor corresponded to concepts used to represent section headers in the terminologies (i.e., metadata) rather than true biomedical concepts (e.g., *General chemical terms*). The reason for excluding these concepts from the comparison is that, due to the polyhierarchical structure of many terminologies, the relation of these concepts to their descendants would often denote a high degree of redundancy, yet not reflect semantic consistency.

RESULTS

We excluded from the results 27,412 relations in which the ancestor corresponds to a section header. The total number of hierarchical relations considered is thus 1,128,261. The total number of hierarchical relations in each family of terminologies is presented in the leftmost column of Table 3.

Proportion of redundant relations

The total number of hierarchical relations present in more than one family of terminologies is 147,939 (13%). The total number of hierarchical relations whose index of redundancy is greater than one is 325,492 (29%). The proportion of relations with an index of redundancy greater than one is presented in Table 3 and in Figure 2 (the two rightmost bars with the \\\ hatching).

Among the 43 families of terminologies studied, the percentage of relations with an index of redundancy greater than one ranged from 0% (PPAC) to 100% (HL7). For one half of the families, the proportion of redundant relations is between 10 and 66%, including Clinical Terms V3 (14%), MeSH (28%), and SNOMED International (37%). Apart from MeSH, the major contributor of hierarchical relations (in absolute numbers) is represented by the Metathesaurus editors who often add the source MTH to existing relations in order to validate them, which makes MTH the first source of redundant relations.

Similarity among terminologies

Similarity among families of terminologies as measured by the cosine method ranged from 0 (for most pairs of families) to .81, with, for example, .79 between SNOMED-2 and SNOMED International. As for SNOMED, pairs with the higher similarity values, shown in Table 1, often consisted of consecutive versions of the same terminology (e.g., ICPC 1 and 2: .70, ICD 9 and 10: .23). However, ICD 9 and 10 exhibit a higher similarity with Clinical Terms V3 (.28, .33) than between themselves (.23). Interestingly, a high similarity value (.66) appears between two unrelated drug terminologies (National Drug Data File and Micromedex DRUGDEX). Not surprisingly, the source MTH representing the work of the Metathesaurus editors exhibits a moderate to high level of similarity with many families of relationships. (Results obtained using the Jaccard method were generally consistent with those presented here).

Families of terminologies		Similarity	Number of redundant relations
HL7	VANDF	0.81	160
SNM-2	SNMI	0.79	12,537
HHC	NAN	0.72	18
ICPC-1	ICPC-2	0.70	373
MMX	NDDF	0.66	3,366
MMSL	MTH	0.63	49,581
DSM-3R	DSM-4	0.50	83
MESH	MTH	0.34	17,351
CTV3	ICD-10	0.33	3,433
AOD	MTH	0.32	12,918
CTV3	ICD-9-CM	0.28	3,433
CSP	MTH	0.27	9,153
CTV3	MTH	0.23	11,757
ICD-10	ICD-9-CM	0.23	1,438
CSP	MESH	0.21	3,020

Table 1 – Similarity among sources (top 15 pairs)

		Semantically consistent		
		Yes	No	Total
Redundant	Yes	314,911	10,581	325,492
	No	771,826	30,943	802,769
	Total	1,086,737	41,524	1,128,261

Table 2 – Redundancy and semantic consistency

Redundancy and semantic consistency

Intuitively, a relation appearing in more than one source is less likely to represent a specific view of the world and is therefore expected to be semantically consistent. This was confirmed by this study, as attested by the results reported in Table 2. However, while 97% of redundant relations are semantically consistent, a similar proportion of non-redundant relations are also semantically consistent. Therefore, the link between redundancy and semantic consistency is weak (Positive Likelihood Ratio = 1.1) and, in practice, redundancy is not sufficient for identifying most of the semantically consistent relations.

With the exception of ICPC-1, the proportion of semantically consistent relations for each family of terminologies ranged from 49 to 100%, the value being over 90% for most of the sources. More details are provided in Figure 2 (the two central bars with the /// hatching).

DISCUSSION

Many studies have been published that investigate specific characteristics of biomedical terminologies, including, for example, content coverage [9] and suitability for natural language processing [10]. This paper sheds light on an aspect of terminologies less frequently studied: hierarchical relations. We now briefly present how the characteristics of hierarchical relations may be exploited in three different kinds of applications.

Selecting sources

Integrating nearly one hundred terminologies, the UMLS Metathesaurus is the most extensive biomedical terminology system, but its volume starts becoming an issue. Moreover, some of its constituent vocabularies have usage restrictions. The characteristics of the hierarchical relations provided by a given source may be a factor in the decision to select a source for an application. For example, as shown in Figure 2, CRISP (CSP) relations are for the most part redundant with relations in other sources and CRISP contributes few specific relations that are semantically consistent; conversely, most relations from the Digital Anatomist (UWDA) are specific and semantically consistent.

Selecting useful relations

We showed that many semantically consistent relations are not redundant and, therefore, redundancy is not a good predictor of semantic consistency. However, semantic consistency is not evenly distributed across the source vocabularies. Factors such as the source and the type of relationship could be used instead to predict semantic consistency. For example, the relationship "narrower than" in the Alcohol and Other Drugs thesaurus (AOD) links to relevant concepts for the purpose of information retrieval. Although useful in this context, the concepts linked by thesaurus relationships are not necessarily expected to be semantically consistent.

Auditing categorization

The link between redundancy and semantic consistency may not be useful for predicting semantic consistency from redundancy, but can be used to audit the semantic categorization of concepts in hierarchical relation. Since redundant hierarchical relations are generally semantically consistent, semantic inconsistency detected in redundant hierarchical relations could be used as an indicator of potential miscategorization of one or both concept, and to trigger a review of these concepts by the Metathesaurus editors. For example, *Stomach Cancer* is a descendant of *Gastrointestinal*, categorized as "Body Part, Organ, or Organ Component", which constitutes a patent semantic inconsistency. The reason here is that the meaning of *Gastrointestinal* in this context is actually something like *Gastrointestinal diseases*, not *Gastrointestinal tract*. Using the lack of redundancy would not have fixed or not even precisely diagnosed the problem. It could, however, be used to draw the attention of the Metathesaurus editors to a potential problem.

In the future, we plan to apply similar methods to concepts shared across terminologies. From the redundancy of both concepts and relations, we will propose a model for a "core" biomedical concept system.

References

1. Chute CG, Cohn SP, Campbell JR. A framework for comprehensive health terminology systems in the United States: development guidelines, criteria for selection, and public policy implications. J Am Med Inform Assoc 1998;5(6):503-10
2. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37(4-5):394-403
3. Solbrig HR, Elkin PL, Ogren PV, Chute CG. A formal approach to integrating synonyms with a reference terminology. Proc AMIA Symp 2000:814-8
4. Spackman KA, Campbell KE. Compositional concept representation using SNOMED: towards further convergence of clinical terminologies. Proc AMIA Symp 1998:740-4
5. Bodenreider O. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. Proc AMIA Symp 2001:57-61
6. Cimino JJ. Auditing the Unified Medical Language System with semantic methods. J Am Med Inform Assoc 1998;5(1):41-51
7. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Medinfo 2001;10(Pt 1):216-20
8. Rasmussen E. Clustering algorithms. In: Frakes WB, Baeza-Yates R, editors. Information retrieval : data structures & algorithms. Englewood Cliffs, N.J.: Prentice Hall; 1992. p. 419-442
9. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classifications. J Am Med Inform Assoc 1996;3(3):224-33
10. McCray AT, Bodenreider O, Malley JD, Browne AC. Evaluating UMLS strings for natural language processing. Proc AMIA Symp 2001:448-52

Family	Number of relations	Redundant	Semantically consistent
AIR	1,103	17.14%	54.31%
ALT	2,449	0.33%	86.97%
AOD	17,732	79.41%	83.10%
BI	1,528	81.87%	48.82%
CCS	980	34.90%	94.59%
CPM	32	31.25%	59.38%
CPT-4	8,838	20.67%	99.67%
CSP	10,978	91.87%	89.58%
CST	2,935	55.06%	91.31%
CTV3	194,094	14.03%	96.52%
DSM-3R	264	67.80%	100.00%
DSM-4	405	54.57%	98.52%
HCPCS	4,019	0.50%	70.91%
HHC	213	18.31%	86.38%
HL7	160	100.00%	70.00%
ICD-10	22,703	26.77%	98.08%
ICD-9-CM	27,264	34.47%	96.74%
ICPC-1	482	82.16%	1.87%
ICPC-2	6,643	26.70%	83.09%
MEDDRA	20,410	30.76%	86.84%
MESH	386,677	28.19%	97.83%
MMSL	62,525	89.56%	98.51%
MMX	11,822	62.17%	96.90%
MST	874	3.66%	88.90%
MTH	233,795	65.53%	96.35%
NAN	149	17.45%	48.99%
NCI	493	16.63%	92.49%
NDDF	18,191	22.67%	95.09%
NEU	808	97.28%	100.00%
NIC	11,057	9.43%	97.87%
NOC	2,620	25.38%	91.30%
OMS	350	3.14%	76.00%
PCDS	1,178	0.08%	94.57%
PDQ	1,809	8.02%	99.12%
PPAC	376	0.00%	97.07%
PSY	5,355	55.00%	82.95%
SNM-2	29,855	69.49%	91.25%
SNMI	101,541	37.48%	97.23%
ULT	56	46.43%	80.36%
UMD	10,512	66.12%	97.45%
UWDA	78,892	10.02%	99.57%
VANDF	11,501	8.56%	94.68%
WHOART	4,661	91.44%	98.56%
Total	*1,128,261*	*28.85%*	*96.32%*

Table 3 – Distribution of the relations

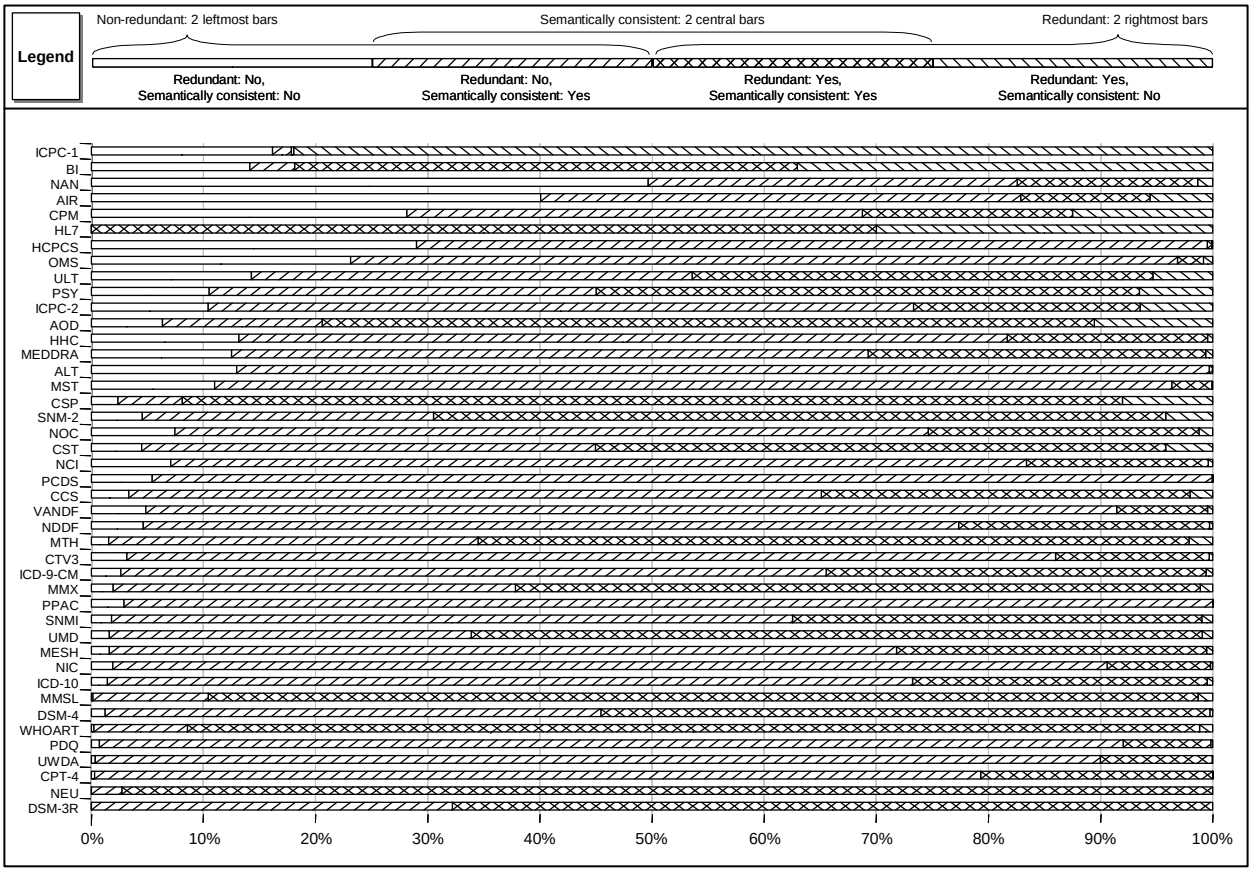


Figure 2 – Redundancy and semantic consistency