Formal Properties of the Metathesaurus®*

Mark S. Tuttle[1], Nels E. Olson[1], Keith E. Campbell, MD[2], David D. Sherertz[1], Stuart J. Nelson, MD[3], William G. Cole, PhD[4]

[1]Lexical Technology, Inc., Alameda, CA; [2]Stanford University, Stanford, CA ; [3]Medical College of Georgia, Augusta, GA; [4]University of Washington, Seattle, WA

The Metathesaurus is a machine-created, human edited and enhanced synthesis of authoritative biomedical terminologies. Its formal properties permit it to be a) exploited by computers, and b) modified and enhanced without compromising that usage. If further constraints were imposed on the existence and identity of Metathesaurus relationships, i.e., if every Metathesaurus concept had a "genus" and a "differentia," then the Metathesaurus could be converted into an "Aristotelian Hierarchy." In this sense, a genus is a concept that classifies another concept, and a differentia is a concept that distinguishes the classified concept from all other concepts in the same class. Since, in principle, these constraints would make the Metathesaurus easier to leverage and maintain computationally, it is interesting to ask to what degree the maintenance and enhancement procedures now in place are producing a Metathesaurus that is also an "Aristotelian Hierarchy." Given a liberal interpretation of the current Metathesaurus schema, the proportion of the Metathesaurus that is "Aristotelian" in each annual version is increasing in spite of dramatic concurrent increases in the number of Metathesaurus concepts.

> *Without formality there is no modifiability nor scalability.[1]*

> *We need formal methods and computer-based tools that can help us with the task [of controlled medical vocabulary construction]. We need research in which controlled vocabulary development is the focus rather than a stepping stone for work on other theories and applications.[2]*

INTRODUCTION

The National Library of Medicine (NLM) Unified Medical Language System® (UMLS®)[3] Metathesaurus is a machine generated and human edited synthesis of authoritative biomedical terminologies that is updated and enhanced annually. Meta-1.0, the first version of the Metathesaurus, was released in 1990, and Meta-1.4, the fifth version, was released in 1994. While the evolution of the form, or *schema*, of the Metathesaurus has slowed, the evolution of *content* has accelerated.

The Schema of the Metathesaurus is Stable

It is the schema of the Metathesaurus that specifies that it is exactly an inter-related set of syntactically homogeneous and semantically unique entries - one entry per concept. Evidence for the current stability of the schema is the fact that the documentation and release format for Meta-1.4, changed only slightly from the documentation and release format for Meta-1.3, continuing a trend begun with the transition from Meta-1.1 to Meta-1.2. A review of the role of "Terminologies," "ATOMS," and "CONCEPTS" in the current Metathesaurus schema appears in **Figure 1**, below.

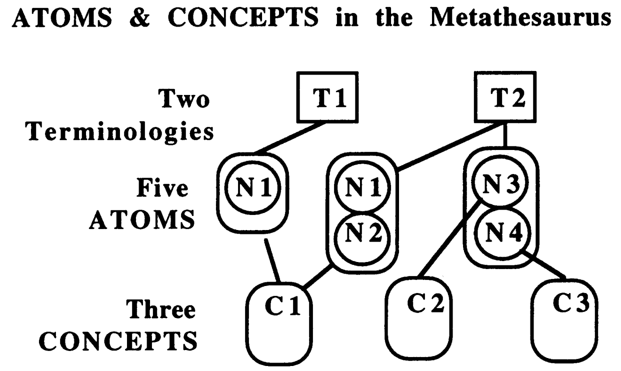

Figure 1 - Two terminologies containing a total of five ATOMS yields three CONCEPTS in the Metathesaurus. Terminology T1 contains a single ATOM, named N1. Terminology T2 contains four ATOMS , namely another instance of N1, and three additional names, N2, N3 and N4. In T2, N2 is a synonym of N1, and N4 is a synonym of N3. From the point of view of the Metathesaurus, T1-N1, and T2-N1,N2 , name CONCEPT C1; N3 names CONCEPT C2; and N4 names CONCEPT C3.

The Content of the Metathesaurus is Increasing

Evidence for the increase in content is displayed in **Figure 2**, below. Since the Metathesaurus grows through the addition of terminologies, one measure of growth is the number of names, or ATOMS, added from each additional terminology. Another measure is the number of new meanings, or CONCEPTS, added to the Metathesaurus by the new names. More precisely, as illustrated in **Figure 1**, we call the occurrence of a name in a terminology an ATOM, and a CONCEPT is an occurrence of a unique named meaning in the Metathesaurus. Thus, both within and between terminologies, more than one ATOM can name the same CONCEPT.

In **Figure 2** the upper line tracks the growth in the number of ATOMS and the lower line the growth in the number of CONCEPTS, in each version of the Metathesaurus. The distance between the two lines represents the degree to which the ATOMS have named the same CONCEPTS.

As implied by **Figure 1**, the Metathesaurus editors are the final arbitrators of whether two ATOMS name the same CONCEPT, or whether they name two different CONCEPTS. On average, the Metathesaurus tends to make fine distinctions, e.g, "Ornithosis" and "Psittacosis" are not synonyms in the Metathesaurus, though they are in some of its constituent terminologies. Further, the Metathesaurus distinguishes “Gentamicins,” a “complex of closely related aminoglycoside sulfates ...,” from “Gentamicin <1>,” a familiar antibiotic, from “Gentamicin <2>,” an assay for the antibiotic. Since, fine or not, these distinctions are maintained only in the “Reviewed” portion of the Metathesaurus, “Unreviewed” entries were not counted. A reviewed entry contains only reviewed ATOMS.[4]

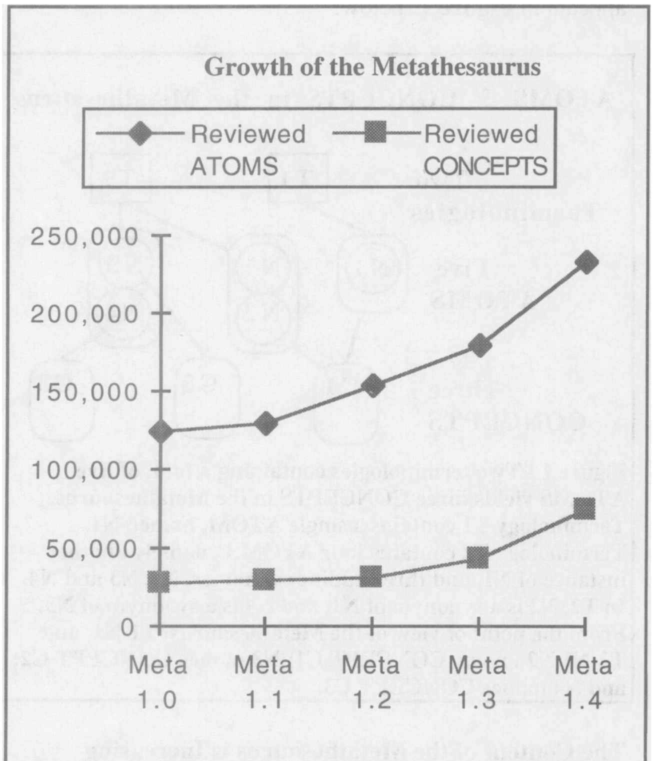

Figure 2 - The number of ATOMS and CONCEPTS in the reviewed portion of the five extant versions of the Metathesaurus: An ATOM is an occurrence of a name in a terminology. A CONCEPT is an occurrence of a unique named meaning in the Metathesaurus. While both ATOMS and CONCEPTS are being added at an accelerating rate, ATOMS are being added faster than CONCEPTS, implying that names are being added faster than meanings.

The Metathesaurus May Be “Converging”

Since, as observed, the Metathesaurus tends to “split” rather than “lump” the concepts named in its constituent terminologies, as with “Ornithosis” and “Psittacosis,” it is remarkable that there is preliminary evidence of “convergence.” That is, for all four annual transitions displayed in **Figure 2**, the rate at which ATOMS were added exceeded the rate at which CONCEPTS were added; and, for the first and third of the three intervals for which it can be computed, the rate of increase of the rate at which ATOMS were added exceeded the rate of increase of the rate at which CONCEPTS were added.

While we believe that the *non-synonymous* relationships between CONCEPTS in the Metathesaurus are what will make the Metathesaurus the most useful in the long run, the practical importance of any potential “convergence” cannot be overestimated. In Meta-1.4 there are 25 terminologies that partially or fully participate in the reviewed portion of the Metathesaurus. Obviously, each ATOM in each terminology was deemed useful by an authoritative body or it wouldn’t have been included. Over the next few years, the number of concepts in the Metathesaurus may double again from the addition of a half-dozen new terminologies alone. If the Metathesaurus continues to show even weak evidence of “convergence” after these additions have been made, then it may mean that there is an “empirical” consensus on what some of the relevant biomedical concepts are, independent of what they are called.

Potential Reasons for “Convergence”

Whether there are such things as intrinsic “concepts” independent of language is a controversy that is more than two millennia old. In brief, the contemporary view can be summed up in two extreme positions. The optimists would assert that any “convergence” of the Metathesaurus would mean that intrinsic “truths” were emerging. The pessimists would assert that we were all just retelling the same “lies,” that is we are all influenced by the same dominant scientific paradigm. Complicating the controversy is the fact that both assertions could be true at the same time, though perhaps in different sub-domains. While the Metathesaurus maintenance and enhancement process represents a unique international experiment, one that may shed new light on this old question, the Metathesaurus is a large extant reflection of “where we’re at,” and it’s hard to imagine any future biomedical terminology efforts ignoring this reality. E.g., even to decide that one wants to do something “different,” is to acknowledge both its existence and its influence. This position is a variation on the notion of “Neurath’s Boat,” (after Otto Neurath), namely, “that we are all at sea without a dry dock; all repairs must be made while we are afloat.”

Accelerating Growth and Its Impact on Developers

Independent of whether or not the Metathesaurus demonstrates a useful degree of convergence, the observation, from **Figure 2**, that the reviewed portion of the Metathesaurus is growing at an accelerating rate is important for developers. Developers will need to decide if their applications that use the Metathesaurus will "scale" to accommodate the new growth.

But what of the complexity, utility and quality of the Metathesaurus? Are these increasing comparably? And, regardless, how will any new complexity, utility and quality affect existing and emerging applications? Metrics for complexity, utility and quality are still being developed for the still immature notion of large-scale, multi-use, terminology enhancement, but one way to begin to assess each of these notions is with respect to an abstract model. One long-standing model is the "Aristotelian" model of classification.

Aristotelian Classification

In the 4th century B.C., the Greek philosopher and polymath Aristotle invented the earliest known classification system for the biological world. This system, employed and much elaborated upon by "Aristotelian" scientists for more than two millennia after his death, served as the foundation for taxonomy until the mid-19th century when Darwin's *Origin of Species* convinced empiricists that they had to take evolutionary relationships into account for proper classification.

> *The standard Aristotelian definition of a form was by genus and differentia. The genus defined the general kind of thing being described; the differentia gave its special character. ... The two together made up the definition, which could be used as a name."[5]*

Linnaeus Rationalized Aristotelian Classification

The 18th century Swedish scientist Linnaeus rationalized the Aristotelian taxonomy by being the first to use binomial Latin nomenclature consistently. Thus, in modern Biology, we have as a member of the genus *dissosteira* (grasshopper) the species *Dissosteira longipennis* (long-winged grasshopper), and from the genus *latrodectus* (spider) the species *Latrodectus mactans* (black widow spider).

In these examples, among many thousands, the Aristotelian classification applied to living things leads to lexical definitions, the differentia, which are incorporated in the names of species. Like these Aristotelian species, concepts in the Metathesaurus can often be seen to have hierarchical relationships that can be interpreted as "genera," and other relationships specifying uniquely defining characteristics that can be interpreted as "differentia."

Genera and Differentia May Support Automation

A potentially important hypothesis is that having "genera" and "differentia" are one way to achieve the computational economies of scale that will be required to sustain the use, maintenance, and enhancement of the Metathesaurus.[6] Thus, even though Linnaean classification suffered from the need to create and understand differentia for larger and larger classes, the hypothesis regarding its compatibility with automation may be true. If it is true, then the extent to which the Metathesaurus is "Aristotelian" is of more than purely historical interest.

Why Formality?

The most important reason to have a Metathesaurus with *formal* properties is to support reproducibility. The schema of the current Metathesaurus[7] is formal in the sense that, in principle, it specifies how ATOMS and CONCEPTS can be added to the Metathesaurus by more than one individual. Further, the current schema lays the foundation for comparable experiments to be done using the Metathesaurus. Those who exploit the schema in the same way should expect comparable results.

The problem with formality is that evolution has not equipped us to deal with it very productively. Humans are "formal," in the sense here, only with difficulty. In addition, whether one believes in the potential power of formality or not, one should keep in mind the problems implicit in the *magnitude* of the numbers appearing in **Figure 2**, and the problems implicit in the scale of the trends to be inferred there, e.g., tradeoffs between formality and tractability per works by C. Cherniak, PhD, on the notion of "undebuggability"[8] and "minimal rationality"[9] and J. Sowa on "local vs. global consistency." However, implicit in the hypothesis that an "Aristotelian" approach is part of the answer is the assumption that human effort will be supported with computational tools.

Naturalistic vs. Experimental Observations

The figures in this paper reflect "naturalistic" observations of the evolution of the Metathesaurus. In brief, there is no notion of "artificially" holding some variables constant while measuring others, as is the case with "experimental observations." Thus, while naturalistic observations do not lend themselves to inferences about causality, they can lead to inferences about correlation.

Why focus on Inter-Concept Relationships?

The remaining results displayed here concern the explicit and implicit relationships between reviewed concepts in the Metathesaurus. In our opinion, these relationships will become the central formal semantics of future versions of the Metathesaurus, independent of the utility of Aristotelian Hierarchies. Relationships will become the dominant representation of meaning because computers can be programmed to manipulate them.

More specifically, one way naming systems specify what their names *mean* is to place those names in a structural context in the naming system. If, as humans, we find these structures semantically impoverished does not mean that they are not useful computationally. In this spirit we explore the past and current state of reviewed inter-concept relationships in the Metathesaurus using a framework adapted, freely, from Aristotle. This framework permitted us to combine years of unilateral and collaborative background study, analysis, and discussion into a single coherent presentation.

THE PROBLEM

Our objective is to determine the degree to which the recent versions of the Metathesaurus represent an "Aristotelian" classification system, given some mappings between the Metathesaurus and "Aristotelian" schemas.

METHODS

As stated, only reviewed concepts, and relationships between reviewed concepts, were analyzed. At present, all *unreviewed* concepts are "Supplementary Chemicals" that are not yet fully Metathesaurus-integrated. Counts were made on the "MR" (Metathesaurus Relational) files. Because of the evolution of the Metathesaurus schema, all counts below were made on the most recent three versions of the Metathesaurus, only. ATOM counts are actually MRSO (Metathesaurus Relational Source) line counts; this ignores a few cases where the same name occurs multiple times in a source without a code. *The fact that Metathesaurus relationships result from separate and combined processes that are themselves axiomatic, lexical, judgmental, principled, and empirical, is ignored.* A more fine-grained analysis would distinguish the *origin* of relationships.

An "exclusive" view of Metathesaurus GENERA would count only "parent" and "broader" relationships as "genera" for a given concept. An "inclusive" view would adds "semantic types" as GENERA, since each type is itself the name of a class in a hierarchy. Thus, since each Metathesaurus concept has one or more semantic types, all Metathesaurus concepts have genera viewed inclusively.

An "exclusive" view of Metathesaurus DIFFERENTIA would count only relationships labeled "other" as a "differentia" for a given concept. While Aristotle's notion of differentia assumes the existence of functions that represent the "essence" of a given form, we assume here that "horizontal" (non-hierarchical) relationships to other concepts are surrogates for such functions. An "inclusive" view would add "definitions," "associated expressions" (ATXs) and "co-occurrences" as DIFFERENTIA, because all could be used by a computer to "differentiate" a Metathesaurus concept from sibling Metathesaurus concepts. Definitions and co-occurrences are assumed to be unique. ATXs do not differentiate concepts unless they are unique, i.e., a few ATXs are identical, currently.

Only the counts for the “inclusive” view of genera and differentia are presented here.

RESULTS

For Meta-1.2, Meta-1.3, and Meta-1.4, and for the "inclusive" definitions, **Figure 3**, below, displays the total number of CONCEPTS - the same data as appears in **Figure 2** - and the number of CONCEPTS with both GENERA and the DIFFERENTIA. The graphs reveal that the degree to which the Metathesaurus is Aristotelian, by our definition, is increasing, though not as fast as the total number of CONCEPTS is increasing.

A refinement of the previous question is to ask it again but only for the 31,064 CONCEPTS common to Meta-1.2, Meta-1.3 and Meta-1.4. That is, for these “sustained” CONCEPTS do the maintenance and enhancement procedures in place increase the degree to which they, alone, are Aristotelian?

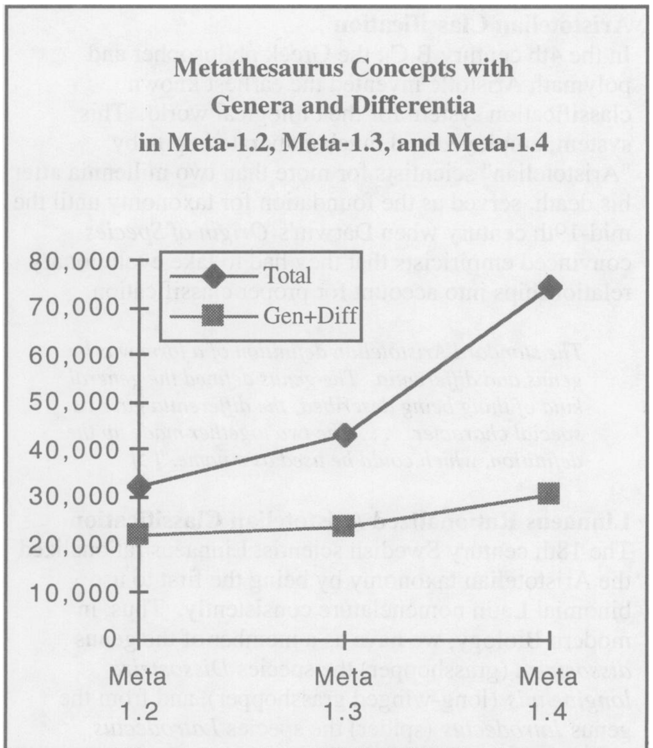

Figure 3 - The total number of CONCEPTS and the number of CONCEPTS with both GENERA and DIFFERENTIA for the last three versions of the Metathesaurus: The number of concepts with both GENERA and DIFFERENTIA is increasing, though not as fast as the total number of CONCEPTS.

Since all Metathesaurus CONCEPTS have GENERA currently, using our definition, the question reduces to one concerning the growth in DIFFERENTIA. For the 31,064 sustained CONCEPTS, 21,383 had DIFFERENTIA in Meta-1.2; 21,763 had DIFFERENTIA in Meta-1.3; and 22,163 had DIFFERENTIA in Meta-1.4. Thus the increases are 380 and 400 additional CONCEPTS with DIFFERENTIA for the two transitions, about 2% per transition. While these increases are small they are potentially significant because they mean that the Metathesaurus maintenance and enhancement process is “naturally” Aristotelian to a small degree, and that part of the observed effect is due to the terminology integration process and not completely to the degree to which the constituent terminologies are Aristotelian already.

DISCUSSION

Examination of the Metathesaurus creation and editing "experience"[10],[11], relative to the Aristotelian notion of classification sharpens three issues: First, the formal needs and cognitive needs to be fulfilled by the Metathesaurus may prove to be different. Second, *when viewed in the aggregate*, any Metathesaurus "persona" to emerge regarding the addition of relationships has yet to dominate, *numerically*, the effect of whatever relationships come with the constituent naming systems.

And, third, Metathesaurus maintenance procedures will have to address the observation that as naming sources are added, Metathesaurus relationships become more tightly entwined.

The first issue brings to mind an early confrontation between cognitive and computational needs. The first time the PDQ (cancer information database) "terms file" was matched against the names in the Metathesaurus, a large number of matches between PDQ names of the form [<body part> <histologic cancer type>], or equivalent, and Metathesaurus names, were put in a report for the physician responsible for review of the PDQ portion of the Metathesaurus "locator" field. When the compound concept did not already exist in the Metathesaurus, the reviewer tended to approve suggested relationships from the "compound" PDQ concept to the "atomic" Metathesaurus concept for the <histologic cancer type>, and tended to disapprove suggested relationships between the "compound" PDQ term and the "atomic" Metathesaurus <body part> concept. While, formally, this seemed like a loss of information, it makes clinical sense. E.g., once a cancer is diagnosed histologically, notions of body site are less important determiners of management and predictors of outcome. This is the "clinical" (human) need, and the anatomic connections would have been less important, and, potentially cluttering cognitively. Of course the relationships might have been useful computationally, independent of their cognitive utility. E.g., combined with other criteria, information about site associations might be used by some future application. Interestingly, however, one "cognitive" technique employed in definitions, namely the appearance of both genera and differentia there, could be exploitable by future automatic methods were it made explicit. For example, the Metathesaurus definition for "Ornithosis," is ...

> *Infection with CHLAMYDIA PSITTACI, transmitted to man by inhalation of dust-borne contaminated nasal secretions or excreta of infected birds. This infection results in a febrile illness characterized by pneumonitis and systemic manifestations.*

An example of the second issue is the critical enhancement of Meta-1.4, namely the mapping of all 18,000 ICD Preferred Terms to MeSH Concepts or MeSH Expressions, so that given a diagnosis, a user can retrieve potentially relevant literature. The magnitude of Metathesaurus growth is now such that this effort, significant by any other measure, is not visible in this analysis.

Relevant to the third issue, one of us (KEC) is developing methods to reduce the "local update penalty."[12] His view is that "Aristotelian compliance" may prove to be an investment that supports coherent maintenance, i.e., before we know whether it would improve the content of the Metathesaurus directly, it will first become necessary computational overhead rather than a cognitive investment in content.

ACKNOWLEDGMENTS

BL Humphreys, WT Hole, MD, PL Schuyler, AT McCray, PhD and RA Zalutsky, PhD, contributed many years of provocative discussions on the notion of "relationships" in the Metathesaurus, and Humphreys and Hole helped clarify some of the specific observations described here. More than a dozen (1986-94) UMLS Contractor Meetings provided feedback on the emerging utility of the Metathesaurus and the relationship of that utility to its formal properties; JJ Cimino, MD, SM Huff, MD, and a sequence of graduate students at Brigham and Women's Hospital, deserve special mention. CG Chute, MD, DrPH, Y Yang, PhD, began analyzing properties of Metathesaurus relationships soon after the first version appeared. MS Erlbaum, MD, LF Fuller, PhD, O Suarez-Munist, SS Lipow and GT Hsu, , LM Fagan, MD, PhD, RD Acuff, MA Musen, MD, PhD, D Oliver, MD, and RW Carlson, MD, contributed directly or indirectly. C Cherniak, PhD, assisted with a review of many of the relevant issues, and J Thomas, PhD contributed to the exposition.

References

*Partially supported by contracts NLM N001-LM-0-3515 and NCI N44-CO-33071. All "®" are held by the NLM.

[1]. Harbison, K, Presentation at the ARPA DSSA Healthcare Workshop, Vail, Colorado, December, 1993.
[2]. Cimino, JJ, "Controlled Medical Vocabularly Construction: Methods from the Canon Group" (Editorial), *J Am Med Informatics Assoc*. 1994; 1:296-7.
[3]. Lindberg, DAB, Humphreys, BL, McCray, AT, "The Unified Medical Language System," *Methods of Information in Medicine*, 1993; Vol. 32, pp. 281-291.
[4]. UMLS Knowledge Sources Documentation, National Library of Medicine, 5th Experimental Edition, 1994.
[5]. Solomon, Arthur K., "Biological Sciences: Taxonomy," *The New Encyclopedia Britannica - 15th Edition*, Chicago, 1992, pp. 965-75.
[6]. Campbell, KE, "Distributed Development of a Logic-Based Controlled Medical Terminology (Dissertation Proposal)," Stanford University, 1994.
[7]. McCray, AT, Nelson, SJ, "The Semantics of the UMLS Knowledge Sources," *Meth Inf Med*, to appear.
[8]. Cherniak, C, "Undebuggability and Cognitive Science," *Comm ACM*, 31, 1988, pp. 402-412.
[9]. Cherniak, C, *Minimal Rationality*, MIT Press, Cambridge, 1986.
[10]. Tuttle, MS, et al., "Implementing Meta-1: The First Version of the UMLS Metathesaurus," LC Kingsland, ed, *SCAMC*, 1989:494-9.
[11]. Sperzel, WD, et al., "Editing the UMLS Metathesaurus: Review & Enhancement of a Computed Knowledge Source," RA Miller ed, *SCAMC*, 90:136-40.
[12]. Tuttle, MS, et al., "Adding Your Terms and Relationships to the UMLS Metathesaurus," PD Clayton ed, *SCAMC*, McGraw-Hill, New York, 1991, pp. 219-23.