Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Biomedical Informatics 36 (2003) 409–413

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Guest Editorial

Research on structural issues of the UMLS—past, present, and future[☆]

Yehoshua Perl and James Geller[*]

Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA

Received 11 November 2003

1. Introduction

This special issue is devoted to research exploring structural issues in the Unified Medical Language System (UMLS). The UMLS was designed by the National Library of Medicine (NLM) to integrate many authoritative biomedical source terminologies into a unified knowledge representation [1–4]. This is an ongoing project involving many external contributors as well as many members of the NLM staff. Careful attention was given by the NLM to conserve the content of each of the original sources, each of which typically represents an authorized list of terms used by a major organization. As a result of these efforts, the UMLS emerged as a huge and complex knowledge base integrating in its 2003AB version more than 100 terminological sources,[1] with a total of 875,255 concepts and 2.14 million terms. For a sample of UMLS research, see the special issue of JAMIA [5] marking the 10th anniversary of the UMLS. For a comprehensive bibliography of papers published during 1986–1996 on UMLS-related work, see [6]. Another bibliography covers the period from 1990 to 2002 [7].

Naturally, most of the research attention in the creation of the UMLS was content-oriented. Tasks such as how to add more source terminologies [8], how to identify common concepts in various sources, how to recognize the terminological sources of a specific concept, relationship, etc. received a lot of attention. However, some design efforts were concentrating on "structural issues," i.e., on mathematical definitions and formal requirements governing the elements of a terminology. Let us illustrate this with a few examples. A directed graph is a mathematical structure. Mandating that the hierarchy of a terminology be a directed acyclic graph (DAG) is a formal (i.e., structural) requirement. A mapping from one set to another is a fundamental mathematical entity. In a terminology hierarchy that has the structure of a tree, the parent relationship is a many-to-one mapping (i.e., a functional mapping). In a DAG, it is a many-to-many mapping because a node may have several parents. The requirement concerning the number of parents of a node is a structural constraint.

A partition of a graph into subgraphs is a mathematical structure, where the structural condition is that each node belongs to one and only one subgraph. Thus, when dealing with structural issues, the emphasis is not on the specific terms of a terminology and what exactly they represent in the real world, but on identifying structures and related requirements to which they must adhere.

Let us see how the previously mentioned mathematical structures relate to the UMLS. A major decision in the design of the UMLS was to have two levels of elements. The Metathesaurus (META) [9,10] contains concepts. The Semantic Network (SN) [11,12] contains general categories, called semantic types. In each of these two levels, there are hierarchical (e.g., IS-A, parent) relationships and semantic relationships connecting pairs of elements. The IS-A hierarchy of the Semantic Network is a structure consisting of two trees that are interconnected by semantic (i.e., non-IS-A) relationships.

An important element of the UMLS is a mapping from META to SN, where each concept is assigned to one or more semantic types. Research concerning the IS-A hierarchy of the SN and the mapping from META to SN is dealing with structural issues. The idea of having a

[☆] This research was supported in part by the New Jersey Commission for Science and Technology.

[*] Corresponding author. Fax: 1-973-596-5777.
E-mail address: geller@la.njit.edu (J. Geller).

[1] http://www.nlm.nih.gov/research/umls/META2.HTML.

1532-0464/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2003.11.005

two-level structure, where the high level Semantic Network helps in the abstraction of and orientation to the large complex META, was pioneering and visionary. To the best of our knowledge, outside of biomedical informatics similar two-level structures did not exist until recently. The Suggested Upper Merged Ontology (SUMO) [13,14], developed towards the IEEE Standard Upper Ontology (SUO), and the mapping of WordNet [15] to SUMO [16] have been conceived in a spirit similar to that of the UMLS two-level structure.

With the UMLS maturing, we have recently seen research that concentrates on structural issues. As a matter of fact, this special issue marks the 15th anniversary of the UMLS. The motivation for this research focus may have been the realization that when dealing with the contents of the UMLS, structural issues in many cases reflect semantic issues. Due to the size and complexity of the UMLS, it is overwhelming to directly approach certain semantic issues. However, the mathematical nature of existing structural constraints enables the design of efficient computational tools, applicable to such a large and complex knowledge base. The results of such computational tools may reflect semantic considerations.

Let us demonstrate structural phenomena in the UMLS, with two examples. In [17], Bodenreider identifies and explores cycles in the IS-A hierarchy of META using graph algorithms. A hierarchy, by definition, may not contain cycles. However, as META contains concepts and hierarchical relationships from many terminology sources, which are not necessarily consistent with each other, such cycles do occur. After identifying the cycles algorithmically, Bodenreider classifies them into various cases according to their semantic causes. This is an example of how structural research can lead to semantic insights.

Another example relates to redundant assignments of concepts of META to semantic types of SN. In the UMLS design, there is a rule that a concept should be explicitly assigned to the lowest (most specialized) possible semantic type in the IS-A hierarchy of the SN [18]. If a concept is assigned to two semantic types A and B such that A is an ancestor of B, then the assignment of the concept to A is called *redundant* [19] and should be removed, according to the above design rule of the UMLS. An algorithm for finding all redundant assignments in the UMLS was given in [20]. Its results can be used to audit and improve the UMLS classifications.

The interplay between the two layers of the UMLS, META and SN, raises some interesting structural research issues that have led to useful results. As the two above examples demonstrate, structural research has helped to expose errors and inconsistencies, unavoidable in an integrated large complex knowledge base such as the UMLS. Such errors would be considerably more difficult to find with purely semantic methods.

The papers which appear in this issue are:

Paper 1: Bodenreider and McCray [21] explore visualization techniques to assess a partition of the SN which they had designed previously [22]. Their techniques are based on visualizing and inspecting for each group of the partition the interactions between the semantic types and their semantic (i.e., non-hierarchical) relationships. This assessment exposes problems in some groups of their partition, while confirming the validity of other groups.

Paper 2: Zhang et al. [23] discuss the design of a metaschema for the Enriched Semantic Network (ESN) [24] of the UMLS. A metaschema [25] is an upper-level abstraction network developed to help with a user's orientation to the ESN. Two metaschemas are derived for two partitions of the ESN; these are the cohesive partition [26] and a modified connected partition of the one originally published in [22].

Paper 3: Cimino et al. [27] explore inconsistencies in the correspondence between the hierarchies of META and SN to expose some errors that may be in either of these two hierarchies or in the assignment of concepts from META to semantic types of SN.

Paper 4: Rindflesch and Fiszman [28] discuss Natural Language Processing of biomedical texts. They show that it is surprisingly difficult to recognize IS-A relationships (to which they refer as hypernymic propositions) from text using only syntactic methods. The UMLS makes it possible to determine whether two concepts are indeed from the same semantic group of concepts and which one of the two is the more general one. Thus, the combination of underspecified syntactic analysis with the UMLS leads to effective processing of natural language sentences in the area of Chemicals and Drugs.

Paper 5: Rosse and Mejino [29] describe the design of the Foundational Model of Anatomy. This design uses a disciplined modeling approach, following a set of developed principles and high-level schemes, applied only in the anatomical structural context rather than in multiple contexts of various application domains. To avoid ambiguity, Aristotelian definitions are used to specify classes in terms of structural attributes.

Paper 6: Mork et al. [30] describe the development of a new query language for semantic networks. This query language is based on an existing declarative language that supports regular expressions. The authors show that their implemented "querying agent" is computationally efficient as well as effective in the digital anatomy domain that they are using. They make convincing claims that their approach will generalize to the complete UMLS.

Taken together, the papers in this special issue constitute a collective contribution. Two papers [Paper 1, Paper 2] present abstraction formalisms for the SN that will help user orientation to the SN. Two other papers [Paper 3, Paper 4] explore the utilization of the mapping of concepts from META to the SN. In the first, the mapping is used for auditing and in the second to support Natural Language Processing of biomedical text. Another two papers [Paper 5, Paper 6] concentrate on creating and using the Fundamental Model of Anatomy terminology intended for extending the anatomy coverage in the Metathesaurus. While extending the UMLS by integration of new terminologies falls under the more traditional content-oriented UMLS research, both these papers are pursuing a structural approach in achieving their goals. In [Paper 5], a structural definition-based approach is described for the design of the terminology. Finally, in [Paper 6], a structural approach is used to design the query language for the terminology.

Assuming that this collection of papers represents a (partial) window into the state of the art of UMLS research, a natural question is: What is the collective message of these papers? The editors of this special issue recognize in all these papers a desire to apply structural (mathematical) techniques to the future development of the UMLS. Harnessing the efficiency of structural techniques and their potential computational applicability to large amounts of data promises results and improvements that would be difficult to achieve by purely semantic techniques, due to their dependence on intensive human labor. Harnessing structural techniques to achieve semantic goals also places this kind of terminology research at the forefront of Knowledge Representation research in general, similar to efforts like the development of the Semantic Web [31–33].

Natural questions that arise are: What should be directions of future development of the UMLS? and What role can structural techniques play in such a development? To address these questions, we review a few issues considered in the early stages of the inception of the UMLS. In [2], Humphreys et al. discuss past differences of opinion between NLM officers and UMLS investigators. Some UMLS investigators suggested building a comprehensive all-inclusive new terminology from scratch, which would have been a major undertaking. However, the NLM decision was to build a compendium of many existing medical knowledge bases in a uniform format instead, allowing user-friendly electronic access. This more limited undertaking was successfully achieved. Looking back, this decision of the NLM was probably the right one. The alternative comprehensive plan might have been too ambitious for its time and the available resources too limited to ensure success.

The NLM's original purpose of integrating many medical terminologies was achieved. Can the UMLS serve as a terminology by itself for the medical field (the alternative plan), although not designed for this purpose? The image that comes to mind is that of a knight, who is so heavily armored that he loses his capability for quick action and mobility. The UMLS, for the purpose of integration, includes all the information that appeared in any of its many sources. But when using the UMLS as a terminology, users typically do not care from which source terminology a concept or relationship was derived. Such a user just wants to know whether there is such a concept or relationship, e.g., whether a given biological agent causes a certain symptom. Anecdotal support for the last claim can be found among the queries that have been sent to the UMLS users' mailing list.[2]

For users who want to know from which source terminology a relationship was derived, the current UMLS will provide the answer. Furthermore, the recent plan for a new format of META, which will provide better "transparency" for the various source terminologies, will further improve this service of the UMLS.[3] On the other hand, the expectation that the current UMLS will serve as a medical terminology by itself is still not justified; the UMLS was not designed for this purpose.

A more realistic plan for designing a comprehensive terminology would be to use the UMLS as a major resource for such an endeavor. What we are proposing, towards this end, is to extract from the present UMLS a lightweight version that would contain all concepts, their terms, attributes and relationships as they appear in the current UMLS. On the other hand, the lightweight version will not include a list of the terminological sources from which a specific concept or relationship was derived. If a pair of terms is linked in various UMLS sources by synonymous relationships, then one of the relationship names will be included in the lightweight version of the UMLS. We would call such a terminology C-UMLS (Core Unified Medical Language System) because it would capture the core knowledge of the UMLS without all the "peripheral" knowledge that is not of interest to many users. The C-UMLS would need to satisfy the desiderata of Cimino [34]. For example, a concept must inherit the relationships of its parent(s). The C-UMLS would have the same two-level structure as the original UMLS, consisting of a Metathesaurus and a Semantic Network.

The C-UMLS would be defined in a way that frees it from the constraints imposed by the need of the UMLS to integrate and preserve all the elements of its terminology sources. For example, the cycles in the IS-A

[2] umls-users@umlsinfo.nlm.nih.gov.

[3] http://www.nlm.nih.gov/research/umls/white_paper.html.

hierarchy of META, detected by Bodenreider [17], as mentioned above, could not be removed from META, as they contain elements of various sources that must be preserved. While building the C-UMLS, such contradictions, which are typically the results of integration between inconsistent sources, would be resolved and no cycles would be allowed to remain in the final concept hierarchy of the C-UMLS. Similar to this example, we expect structural techniques to play a major role in a project of creating a C-UMLS terminology.

In the following, we will mention several issues where structural methods, similar to those reported in this issue, can contribute to the development of the C-UMLS. Correspondence of relationships can be extended from hierarchical relationships [Paper 3] to semantic relationships. That is, if there is a relationship REL from a semantic type A to a semantic type B in SN, then there should be a corresponding relationship REL from a concept assigned to A to a concept assigned to B. In the current META, such a relationship may exist with the name REL, or may be missing, or may exist with another name. Enforcing such a correspondence will help to improve and complete the coverage of relationships in META. This will also improve the support for Natural Language Processing for medical texts [Paper 4].

Systematic inheritance of relationships from a concept of META to all its children will enable the application of partitioning [Paper 1], [22] and schema abstraction [Paper 2], [25] to META, as opposed to the Semantic Network, as done in the above papers. Such partitioning should utilize the structure of the concepts, which can be defined for each concept as the set of its relationships, similar to what was done in [35] for the Medical Entities Dictionary [36]. Applying such partitioning to the (mostly large) sets of concepts of a given semantic type will help auditing and orientation to them. Due to the size of META, partitioning to support orientation to its elements is more critical than for SN, as is done in [Paper 1, Paper 2].

To accelerate the integration of a new terminology into the UMLS, such as of the FMA terminology [Paper 5], one should create a classification of its concepts by the semantic types of SN [37]. A systematically structured META will enable the application of efficient query languages, as the one developed in [Paper 6]. These ideas demonstrate that, although the creation of a C-UMLS would take a substantial effort, techniques to support such a project exist and more are expected to emerge.

In summary, structural techniques can help with efficiently maintaining, auditing and using the UMLS as well as other terminologies. Furthermore, research on structural issues in terminology representations provides a framework for abstraction of terminologies, which can be used as the basis for gaining an orientation to very large and complex terminologies.

Acknowledgments

J. Cimino was instrumental in developing this special issue. We thank him for his advice along the way. We thank T. Shortliffe for his continuous guidance and patience with us. We thank all the anonymous reviewers for their contributions that had a major impact on the quality of the papers.

References

[1] Campbell KE, Oliver DE, Shortliffe EH. The Unified Medical Language System: toward a collaborative approach for solving terminologic problems. JAMIA 1998;5(1):12–6.

[2] Humphreys BL, Lindberg DAB, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. JAMIA 1998;5(1):1–11.

[3] Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993; 32:281–91.

[4] Humphreys BL, Lindberg DAB. Building the Unified Medical Language System. In: Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care, November 1989; Washington, DC. p. 475–80.

[5] McCray AT, Miller RA (Editors). Making the conceptual connections: The UMLS after a decade of research and development. Special Issue of the Journal of the American Medical Informatics Association, 5(1), January/February 1998.

[6] Seldon CR, Humphreys BL. Unified Medical Language System. Curr Bibliographies Med 1997;8:96.

[7] Unified Medical Language System Bibliography. US Department of Health and Human Services, NIH, National Library of Medicine. Available from: http://umlsinfo.nlm.nih.gov/bibliography.html.

[8] Tringali M, Hole WT, Srinivasan S. Integration of a standard gastrointestinal endoscopy terminology in the UMLS Metathesaurus. In: Proceedings of the 2002 AMIA Annual Symposium, November 2002; San Antonio, TX. p. 801–5.

[9] Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. The UMLS Metathesaurus: representing different views of biomedical concepts. Bull Med Libr Assoc 1993;81(2):217–22.

[10] Tuttle MS, Sherertz DD, Olson NE, et al. Using META-1, the first version of the UMLS Metathesaurus. In: Proceedings of the Fourteenth Annual SCAMC, 1990. p. 131–5.

[11] McCray AT. Representing biomedical knowledge in the UMLS Semantic Network. In: Broering NC, editor. High-Performance medical libraries: advances in information management for the virtual era. Westport CT: Mekler; 1993. p. 45–55.

[12] McCray AT, Hole WT. The scope and structure of the first version of the UMLS Semantic Network. In: Proceedings of the Fourteenth Annual SCAMC, November 1990; Los Alamitos, CA. p. 126–30.

[13] Niles I, Pease A. Origins of the IEEE Standard Upper Ontology. In: Working Notes of the IJCAI-2001 Workshop on the IEEE Standard Upper Ontology, Seattle, WA, August 2001.

[14] Niles I, Pease A. Towards a Standard Upper Ontology. In: Proceedings of the FOIS 2001, Ogunquit, MA, October 2001.

[15] Fellbaum C. WordNet: an electronic lexical database. Cambridge, MA: The MIT Press; 1998.

[16] Niles I, Pease A. Linking lexicons and ontologies: mapping WordNet to the Suggested Upper Merged Ontology. In: Proceedings of the International Conference on Information and Knowledge Engineering 2003 (IKE'03) p. 412–6.

[17] Bodenreider O. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. Proc AMIA Symp 2001;23:57–61.

[18] McCray AT, Nelson SJ. The representation of meaning in the UMLS. Methods Inf Med 1995;34:193–201.

[19] Gu H, Perl Y, Geller J, Halper M, Liu LM, Cimino JJ. Representing the UMLS as an object-oriented database: modeling issues and advantages. JAMIA 2000;7(1):66–80.

[20] Peng Y, Halper M, Perl Y, Geller J. Auditing the UMLS for redundant classifications. In: Proceedings of the 2002 AMIA Annual Symposium, November 2002; San Antonio, TX. p. 612–6.

[21] Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. J Biomed Inform 2003;36:414–32.

[22] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. In: Proceedings of the Medinfo 2001, September 2001; London, UK. p. 171–5.

[23] Zhang L, Perl Y, Halper M, Geller J. Designing metaschemas for the UMLS Enriched Semantic Network. J Biomed Inform 2003;36:433–49.

[24] Zhang L, Perl Y, Geller J, Halper M, Cimino JJ. Enriching the structure of the UMLS Semantic Network. In: Proceedings of the 2002 AMIA Annual Symposium, November 2002. San Antonio, TX. p. 939–43.

[25] Perl Y, Chen Z, Halper M, Geller J, Zhang L, Peng Y. The cohesive metaschema: a higher-level abstraction of the UMLS Semantic Network. J Biomed Inform 2003;35(3):194–212.

[26] Chen Z, Perl Y, Halper M, Geller J, Gu H. Partitioning the UMLS semantic network. IEEE Trans Inf Technol Biomed 2002;6(2):102–8.

[27] Cimino JJ, Min H, Perl Y. Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. J Biomed Inform 2003;36:450–61.

[28] Rindflesch TC, Fiszman M. The interaction of dynamic knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform 2003;36:462–77.

[29] Rosse C, Mejino LV. A reference ontology for bioinformatics: the Foundational Model of Anatomy. J Biomed Inform 2003;36:478–500.

[30] Mork P, Brinkley JF, Rosse C. OQAFMA querying agent for the foundational model of anatomy: a prototype for providing flexible and efficient access to large semantic networks. J Biomed Inform 2003;36:501–17.

[31] Berners-Lee T, Hendler J, Lassila O. The Semantic Web. Sci Am 2001;284(5):34–43.

[32] Fensel D, Horrocks I, van Harmelen F, McGuinness D, Patel-Schneider PF. OIL: ontology infrastructure to enable the Semantic Web. IEEE Intell Syst (Special Issue on the Semantic Web) 2001;16(2).

[33] McIlraith S, Son T, Zeng H. Semantic web services. IEEE Intell Syst (Special Issue on the Semantic Web) 2001;16(2):46–53.

[34] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inform Med 1998;37:394–403.

[35] Gu H, Halper M, Geller J, Perl Y. Benefits of an object-oriented database representation for controlled medical terminologies. JAMIA 1999;6(4):283–303.

[36] Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. JAMIA 1994;1(1):35–50.

[37] Lee Y, Supekar K, Geller J, Perl Y. Using algorithmic semantic refinement for ontology integration. 2003 [submitted].