A Role for Controlled Vocabularies in Developing Structures for Sharing Medical Knowledge

Carol A. Bean, Ph.D.

National Library of Medicine
Lister Hill National Center for Biomedical Communications
8600 Rockville Pike
Bethesda, MD 20894
bean@nlm.nih.gov

Abstract

Controlled Medical Vocabularies (CMVs) have been studied to identify and characterize organized sets of well-specified semantic relationships in the biomedical domain. Results of research on both hierarchical and non-hierarchical associative relationships in existing CMVs are being used to enrich ontological content through the discovery of subclasses, attributes, and relationships, and to enhance structure by facilitating the organization of domain knowledge and exploiting relational patterns in their hierarchies and other forms.

INTRODUCTION

Our conceptual world is most commonly and effectively modeled by representing its component entities and the relationships that obtain among them. However, relatively little attention has been directed to the development of a rich set of well-defined relationships for the organization of knowledge in medicine, especially in contrast to the vast set of well-defined concepts that exists. This paper describes some issues in empirical relationship discovery and the role these relationships might play in the development and integration of knowledge organization structures. These studies have largely focused on the subject domain of anatomy, and where necessary, been constrained to locative (physical and spatial) relationships.

These studies addressed the following issues: What are the predominant relationships and classes? Do semantic patterns or clusters exist for both relationships and concepts? Do these patterns vary by conceptual context, or among different subject categories or trees? What relationships exist between the relationships? To what extent is each tree in the overall hierarchy itself mixed or pure; i.e., are the relationships the same in a given branch of the tree? What patterns are displayed in individual trees as compared to those seen in semantic clusters or groupings? How consistent are the patterns within a particular subdomain? Across subdomains?

CONTROLLED VOCABULARIES AS DOMAIN KNOWLEDGE BASES

A source of extensive structured domain knowledge is necessary for a variety of tasks. A domain model satisfies this need by defining the entities and relationships in some world. Characteristics of the best domain models include extensive breadth of coverage, relationships explicitly encoded as rules, and its entities are atomic concepts (or where complex, the internal relationships are explicitly defined) (Rindflesch et al. Forthcoming). Domain models as we know them typically represent but a single perspective on a particular (single implied) domain. Some of the problems in integrating or sharing knowledge from disparate systems stem from contextual idiosyncracies among the various underlying conceptual models, even in closely aligned subject domains. There exists a need for better characterizations of individual domains as well as some sort of Super- or Meta-Model to represent multiple perspectives on a single domain.

Controlled vocabularies are (underspecified) knowledge bases that provide one or more perspectives on a given subject domain from a particular point of view; in other words, a domain model. Vocabulary content is determined by the subject domain, and the organization of that content reflects a particular perspective on that domain, i.e. context. (The exact subset of domain knowledge represented is also determined to some degree by the perspective.) Much of the knowledge in a CMV is contained within its syndetic, or relational, structure. The syndetic structure of a controlled vocabulary may then be seen as an organized expression of the relationships (equivalence, hierarchical, and associative) among its concepts, and used to discover implicit and to characterize explicit relationships. Precise specification of the myriad vocabulary structures in a domain will provide a contextual dimensionality for the knowledge contained in each that is sufficient to support their integration.

The efforts to derive a common description that incorporates and integrates a myriad of perspectives may be compared to the old story about three blind men standing around an elephant trying to describe that elephant from what they can feel: Each has a different perspective on the same subject. Like them, we would benefit from a shared perspective that derives from and incorporates the elements of each individual perspective. Thus, it may be useful to distinguish between various models of the knowledge in some domain and a (composite) model of the domain itself.

STUDY RESOURCES

The National Library of Medine's Unified Medical Language System (UMLS) has provided the data resources for the studies of semantic relationships in CMVs. Designed primarily for use by system developers, the UMLS Knowledge Sources comprise four components: Metathesaurus, Semantic Network, SPECIALIST Lexicon, and Information sources Map. The Metathesaurus is a compendium of vocabularies that are conceptually linked by relationships from the Semantic Network, while at the same time each retains its local relationships. Each concept in the Metathesaurus is assigned to one or more general categories or Semantic Types in the Semantic Network. In the 1998 release (9th edition), the Metathesaurus contains almost half a million concepts and over a million different concept names from more than 40 vocabularies, classifications, and coding systems. The Medical Subject Headings (MeSH) is the best-known and most widely used controlled medical vocabulary, and is an important component of the UMLS. Originally designed for and primarily used to serve bibliographic information retrieval needs, MeSH is reviewed and revised annually. The current Semantic Network has 132 types and 53 relationships.

UMLS data files are distributed to licensees under experimental agreement on CD-ROMs or by ftp. Access to UMLS Knowledge Sources is available over the Internet via the NLM's Unix-based Knowledge Source Server via Web-based browser, command-line, and API interfaces. Recent evidence indicates that the combination of controlled vocabularies contained within the UMLS provides adequate coverage of the majority of terms needed to describe patient conditions (Humphreys, McCray & Cheh 1997), and so provides both raw data and tools (McCray et al. 1996) for further analyses of needs for knowledge integration in the clinical domain. While each of the 41 component vocabularies in the Metathesaurus retains its own conceptual context, the Semantic Net provides an overarching integrative context within the biomedical domain at large.

HIERARCHICAL ASSOCIATIVE RELATIONSHIPS

Survey of Hierarchical Relationships

In a given information system, the exact meaning of a concept is determined by the context in which it occurs; the relationships a concept has with other concepts in the system will define its context and thus its meaning. Contextual information in knowledge structures is most often conveyed via hierarchy. What principles the hierarchy and its subunits are organized around may be seen to reflect the predominant organizing principles of the domain itself. Hierarchy has long been the dominant structuring mechanism in knowledge organization. There have been numerous efforts to inventory hierarchical relationships. While the resulting lists vary somewhat, most investigators would agree on the primacy and predominance of two hierarchical relationships. The most common, and perhaps prototypical, hierarchical relationship is Is-a, which describes the relationship between a class and a subclass or a type and its instantiation. The other primary hierarchical relationship is Part-of, which most typically describes aggregation or composition. Vocabulary and knowledge-base developers do not always distinguish between hyponymy (Is-a) and meronymy (Part-of) in their type hierarchies, often mixing them both among and within individual trees.

Bean (Forthcoming) examined the nature of explicitly specified hierarchical relationships found in different subject contexts, i.e. MeSH tree structures. The MeSH conceptual organization is a poly-hierarchical taxonomy, having 15 broad subject categories with 102 narrower subcategories called "trees." While the primary hierarchical relationship is understood to be Is-a, MeSH trees, like so many other vocabularies, contain a mixture of relationships. Although Is-a relationships were predominant overall, over a third of specified Parent-Child relationships in MeSH98 were other than Is-a, comprising an additional 67 different relationship types. Over 80 percent of both Parents and Children fell into 3 of 15 main Semantic Type Groupings (high-level class clusters). Likewise, five different Semantic Net relationships accounted for almost two-thirds of all explicit non-Isa Parent-Child relationships.

Domain Relational Profiles

Because the conceptual content of a controlled vocabulary covers an expressly delimited subject domain, the relationships among the concepts might be expected to exhibit parallel patterns of domain dependency. Recent evidence seems to support the existence of a set of general relationships (e.g., Part-of, Causes) that would apply to almost all subject domains (Bean & Green 1997; Green 1997). A specialized subject domain such as Medicine

could exhibit domain specificity by supplementing this shared general set with additional more specialized relationships (e.g., Diagnoses, Treats). Another way that domains could be distinguished through their relationships might be by asserting the general relationships among a different set of concept classes or in a novel pattern. In this way, the relative proportion or frequency of particular relationships would be expected to vary among subject domains, presenting a different, characteristic pattern or profile for each. These patterns should be most evident in the syndetic structure of the controlled vocabularies of a constrained subject domain. Semantic relationship patterns can be examined from both horizontal and vertical perspectives to elucidate hierarchical and associative clusters, respectively. (In most vocabularies these also correspond to formal vs informal structural elements.)

In this study (Bean, Forthcoming), all Parent-Child relationships were tabulated for 10 of the 16 Anatomy and 2 of the 26 Disease trees, and classified according to broad Semantic Net Relationship groupings. (The trees chosen provided a sample representative of the semantic clusters of high-level anatomic entities in a symbolic knowledge model of anatomy derived previously. (See Bean et al. 1996.)) The relational profiles of these MeSH trees were found to vary both within and among subject subdomains, but tended to display characteristic domain patterns. Several striking patterns emerged that served to distinguish not only between the broad categories of Anatomy and Diseases, but also among the broad groups of Anatomic classes.

In general, the Anatomy groupings were largely characterized by Identity, Spatial, and Physical relationships, which accounted for more than 90 percent of these Parent-Child relationships. The Disease trees had a large proportion of Identity Parent-Child relationships (83 percent); however, the distinguishing feature of this profile was the presence of numerous Functional, and to a lesser degree, Conceptual relationships (11 and 6 percent, respectively). In contrast to the Anatomy groups, no Physical or Spatial relationships were present among these Disease Parent-Child relationships; on the other hand, there was only a single Functional relationship among all 764 Parent-Child relationships examined for the Anatomy trees. About half of the Parent-Child relationships in Anatomic Spaces (Body Regions) were Spatial, which were not nearly as common in any other groups. A variety of patterns was seen for the Anatomic Structures/Systems groups, probably reflecting the functional organization of these trees and the tendency for some systems to show a greater correlation between structure and function than others. However, these groups were most striking for the relative frequency of Physical relationships, and for the frequent presence of Conceptual relationships in the trees. Identity (Is-a) relationships predominated in Anatomic Substances trees (80 percent).

NON-HIERARCHICAL ASSOCIATIVE RELATIONSHIPS

To the extent that the relationships among concepts are explicitly expressed in a controlled vocabulary, most focus on hierarchical relationships. Far less attention is paid to non-hierarchical, or associative relationships. Where they are identified, they are rarely characterized beyond being grouped with other similarly non-specified relationships under such descriptive designations as "related terms," "see also," "see related," or "other." The larger vocabularies may contain hundreds or thousands of these relationships. Although the existence of such implicit hierarchical associative relationships has long been formally recognized by explicit reference, this is of little practical use when the precise meaning or significance of the relationship between the concepts or terms in the context of the vocabulary is unknown because it is only implied. While the nature of the relationship may be or become apparent to subject experts or experienced users, novices to either the content area or the workings of the particular information system are often at a loss to discern meaning. Two studies have examined non-hierarchical associative relationships in MeSH (Bean 1996) and in other CMVs in the UMLS (Bean 1997), identifying semantic clusters based on "See Related" relationships within pairs of Focal Terms and their associated Related Terms.

In the MeSH study, 256 such pairs whose Focal Term originated from the gross anatomy trees were clustered into four broad semantic categories that reflected both the linked terms and the nature of the relationships between the terms. The most frequent category was Procedures, where over half the Related Terms clustered, followed equally by other Anatomic Entities and Functions, with a few Chemical Agents. All of the Procedure relationship pairs linked specific procedures and the anatomic entities they were performed on. The relationships among Anatomic Entities described spatial or compositional links between terms. Using a somewhat different methodology that sampled more than 1,700 additional See-Related term pairs on the basis of focal-term semantic type, the UMLS study yielded very similar results, with almost identical distribution patterns. Different subject domains showed the same range of relationship types, but in different proportions, in a manner consistent with the intended purposes of the vocabularies.

The presence of such semantic clusters supported the notion of "families" of related relationships, but both studies noted a need for finer conceptual granularity to support the expression of additional relationships among subclasses, as well as the need to express n-ary relationships, with more than two arguments. Thus, ample empirical evidence exists to suggest the presence of characteristic patterns of horizontal and vertical relationships for subdomains that may be considered as

relational profiles.

Results from the UMLS study suggested several categories of potential changes that might improve the utility of the Semantic Network as a domain model. First, the need to establish formal links among existing classes became apparent; for example: analysis of the UMLS "other" relationships revealed numerous Related-Term relationships between concepts of the Semantic Types THERAPEUTIC OR PREVENTIVE PROCEDURE and BODY PART, ORGAN, OR ORGAN COMPONENT when no legal Semantic Net relationships existed for this pairing. A second type of refinement suggested the specification of subclasses for existing classes on either side of the relationship. Current UMLS Semantic Types distinguish among procedures on the basis of motivation. The Related-Term clusters suggested classes based on the specific actions taken on the Focal Term, for example removal, ablation, or excision. Likewise, Focal Term clusters also suggested subclassing concepts based on characteristics of anatomic form. For example, an '-ostomy' procedure (which creates an opening in a structure) would only be performed on Body Parts that are hollow or tubular, such as the hollow viscera.

DISCUSSION

Conceptual Model Building

Knowledge structures are far more than the sum of their concepts. Their optimal construction and use in operation, as well as their integration across disparate systems and the sharing of knowledge therein, requires an explicit understanding of the organizational principles underlying their structure. The real significance of this line of research to knowledge organization and integration may lie in its attempts to develop a principled methodology for empirical relationship discovery and establish an.empirical basis for design decisions in conceptual modelling of domain-specific knowledge. Knowledge sources themselves then contain at least some of the keys to integrating them one with another via their syndetic structure.

Application of these findings to knowledge model development provides an orthogonal approach to traditional methods of conceptual model design, which rely on bottom-up methods of instantiation to flesh out an overall structure somewhat intuitively derived from a top-down perspective. The expected advantage derives from the incorporation of real data as part of the building process, which should yield a more realistic model of the conceptual world, more consistent with the real world. This would presumably require less change or adjustment during the population (or instantiation) phase of model development, and during maintenance, thus extending the useful life of the knowledge model. Relationships in knowledge structures also help keep class proliferation in check, serving as a sort of functional Occam's razor. Increasing the explicit representation of relationships provides a richer more informative ontology, which should aid alignment of different ontological structures. In short, enhancing the relational structure in a knowledge model makes integration more efficient, more effective, and more reliable, and thus facilitates knowledge sharing.

Operations on Knowledge Structures

Contextual information in knowledge structures is specified by explicit and implicit relationships. The principles used to organize a knowledge structure may be seen to reflect the predominant organizing principles of the subject domain itself. While the investigation of syndetic structural knowledge in controlled vocabularies may be seen as both a means to end and a worthy end unto itself, this paper has focused on the former, that is, on controlled vocabularies as a set of resources containing both the content and structure needed to build robust, multipurpose (i.e., integrative) knowledge models. This line of research demonstrates the necessity to make explicit all interconcept links in a controlled vocabulary, even the hierarchical ones, if we are to be able to exploit their inherent syndetic structure. An increased awareness and understanding of these relationships will inform our reliance on certain assumptions underlying basic principles of organization in knowledge structures.

Various aspects of the logic of hierarchical relationships bear further investigation. Operations on hierarchies depend on several assumptions about the relationships they are structured around, and on three in particular. These are directionality (typically expressed as some sort of superordination), inheritance, and transitivity. Because hierarchy implies some sort of precedence or governance of one participant in the relationship over the other, each relationship asserted in a hierarchy can be seen to have an inherent direction. One expects that hierarchical relationships, which typically characterize superclass-subclass or type-token relationships, will "go" in one particular way, with their reciprocal or inverse relationship corresponding to the reverse direction. The principles of transitivity and inheritance are the essential hallmarks of hierarchical knowledge structures, and are used extensively in operations on them. These properties make hierarchical taxonomy both a cognitively satisfying and a computationally powerful structure for organizing knowledge. The economy and efficiency they allow have made it the standard structure for knowledge representation. However, these properties reliably apply only to the Is-a relationship. That these principles would be affected by mixed-relationship hierarchies and by non-hierarchical structures is obvious, but precisely how, and more importantly, how they might be used to improve retrieval on semantic propositions, remains to be seen. Meronymic transitivity is not as well understood, most likely because of the myriad types of Part-of relationships; neither has the nature of transitivity been determined for

the variety of other relationships that might be present in knowledge structures.

Our computational use of hierarchical knowledge structures relies on these properties, which vary among different relationship types. Further, it is clear that many Parent-Child relationships in the medical domain are not prototypically hierarchical after all, and these principles can not necessarily be assumed to hold. It is necessary to identify what relationship types actually do exist in hierarchies, and then to determine the operating logic of such relationships. Because classes may be distinguished by the relationships they enter into, domain-specific knowledge structures can be organized to reflect and exploit this.

REFERENCES

Bean CA. The Semantics of Hierarchy: Explicit Parent-child Relationships in MeSH Tree Structures. (Forthcoming) IN Mustafa-ElHadi W. (Ed) *Structures and Relations in Knowledge Organization. Proceedings of the Fifth International ISKO Conference*, Lille, France, August 25-29 1998.

Bean CA. 1996. Analysis of Non-hierarchical Associative Relationships among Medical Subject Headings (MeSH): Anatomic and Related Terminology. (1996) IN Green R (Ed) *Knowledge Organization and Change. Proceedings of the Fourth International ISKO Conference,* Washington DC, USA, July 15-18, 1996. Frankfurt/Main:Indeks Verlag. pp.80-6.

Bean CA. 1997. Development of a Knowledge Model of Clinical Anatomy Based on Semantic Relationships Empirically Derived from Controlled Medical Vocabularies. (1997a) Paper presented at Annual Meeting of the Medical Library Association, Seattle WA, May 24-27, 1997.

Bean CA, Green R. 1997. Development of a Structured Inventory of Relationships. Presented at ACM-SIGIR Workshop: *Beyond Word Relations*. Philadelphia PA, July 1997.

Bean CA, Molholt P, Imielinska C, Laino-Pepper L. 1996. Symbolic and Spatial Knowledge Model. In Banvard R (Ed) *The Visible Human Conference Proceedings,* Bethesda MD, October 7-8 1996.

Green R. 1997. A Relational Thesaurus; Modeling Semantic Relationships Using Frames. *Final Report to OCLC Online Computer Library Center.*

Humphreys BL, McCray AT, Cheh ML. 1997. Evaluating the Coverage of Controlled Health Data Terminologies: Report on the Results of the NLM/AHCPR Large Scale Vocabulary Test. *Journal of the American Medical Informatics Association* 4(6):484-500.

McCray AT, Razi AM, Bangalore AK, Browne AC, Stavri PZ. 1996. The UMLS Knowledge Sources Server: A Versatile Internet Based Research Tool. Cimino JJ (Ed) *Proceedings fo the 1996 AMIA Annual Fall Symposium.* Philadelphia:Hanley and Belfus, 1996, 164-8.

Rindflesch T, Aronson A, Hole WT. Forthcoming. The UMLS as a Domain Model of Medicine. Manuscript submitted to the 1998 AMIA Fall Symposium.