Construction of a Medical Informatics Thesaurus

Nancy J. Ogg[*+], MaryEllen Sievert[*], Zong Rong Li[^], Joyce A. Mitchell[+]

[*]Dept. of Information Science, University of Missouri-Columbia
[+]Medical Informatics Group, University of Missouri-Columbia
[^]Medical Informatics Institute, Hubei Medical University, China

ABSTRACT

Medical Informatics needs a specific terminology that reflects the multi-disciplinary nature of the field and can rapidly change as the discipline evolves. Using the four primary methods of thesaurus construction, a group at the University of Missouri-Columbia is developing a thesaurus that meets these criteria and includes both user and literary warrant. The steps in construction include using existing thesauri, medical informatics literature, and the terminology of experts to produce a thesaurus arranged within a hierarchical structure.

INTRODUCTION

Medical Informatics is a constantly evolving, multi-disciplinary field which draws terminology from a variety of disciplines. When we tried to do original abstracting and indexing of medical informatics literature for the Medical Informatics Information Center at the University of Missouri-Columbia we realized a specific informatics thesaurus was needed. Several steps were taken to confirm that this need did exist and to identify its range.

First, we reviewed the history of the L Tree of MeSH[1] because it was the first vocabulary to include medical informatics terminology and is still used as a source of medical informatics terminology. We found that the current L Tree terms were not sufficient to cover the discipline because there were numerous major concepts not covered. An examination of the thesauri used in the contributing disciplines and of the medical informatics literature both confirmed our view and allowed us to develop a raw vocabulary of medical informatics terminology now in use. The resulting thesaurus will contain a manageable number of terms with an accompanying classification.

We believe that the medical informatics thesaurus we are developing more adequately covers the existing medical informatics discipline than anything currently does. It can also be quickly revised to include the rapidly evolving terminology of the discipline. The purpose of this paper is to present the steps we used to create this new thesaurus and to discuss our plans for further testing, refinement, and publication of that thesaurus.

METHODS

Our purpose in developing the thesaurus was to address both user and literary warrant in selecting terms for inclusion. User warrant means that the terms chosen for inclusion must be those which users in the field would use, while literary warrant means that the terms would be found in key documents in the literature of that discipline.

In order to achieve this goal, we looked at the four primary methods of thesaurus construction and used parts of each of them for our project. The four main methods of thesaurus construction, as identified by Lancaster[2], are to:

1. Generate the vocabulary empirically on the basis of indexing a representative set of documents;

2. Convert an existing vocabulary;

3. Extract the vocabulary from an existing, more general thesaurus or develop a specialized thesaurus within the framework of a general one;

4. Collect terms from diverse sources including glossaries, other publications, and from subject specialists.

The following discussion shows how we included facets of each of these approaches into our thesaurus construction.

0195-4210/94/$5.00

Work on this project began in September 1992. A historical review of the MeSH L Tree was performed to ascertain the degree and nature of the changes made in the terminology in this source over the last thirty-three years.[3] A search of fifty titles from papers in the 1992 Symposium on Computer Applications in Medical Care (SCAMC) identified terms used in these titles that also appeared in the 1992 L Tree. Major concepts in the 1992 SCAMC for which no terms appeared in the L Tree were also identified. Examples of such concepts include Bayesian networks, knowledge bases and knowledge representation, heuristic approaches, electronic patient records, physician order entry, nursing systems and nursing informatics. These findings indicate that the L Tree would not suffice as a thesaurus for the medical informatics literature.

Because of the multi-disciplinary nature of the medical informatics field and the lack of a comprehensive thesaurus, seven thesauri from other fields were examined for relevant terminology for indexing medical informatics articles. The first step in the thesaurus construction was to compare the terminology in each of these thesauri to see which terms were used by all, some or only one of them. The thesauri used and the fields represented include:

Engineering field:

Engineering Information Thesaurus (EI)[4]
INSPEC Thesaurus[5]

Education/psychology field:

Thesaurus of ERIC Descriptors[6]
Thesaurus of Psychological Index Terms[7]

Medical field:

Medical Subject Headings (MeSH)[8]

Computer field:

Computer Select® Glossary of Terms[9]

Informatics field:

In-house list of indexing terms used at the Medical Informatics Information Center at the University of Missouri-Columbia.

Terms from each thesaurus were assigned a weight according to their importance in the informatics field. These weights were assigned to allow us to determine which terms were most significant to the literature. A term appearing in both the medical and the engineering terminology should be considered more important than a term carried only in the psychology terminology. Thesauri from the fields of engineering (EI and INSPEC) and medicine (MeSH) received a weight of two. The rest were assigned weights of one. The total weight for a particular term was determined by how many and which thesauri included the term.

The second step was to study the medical informatics literature to determine what terminology those who work in the field use. The literature examined came from journals and proceedings. The following journals were considered key medical informatics journals:

1. "Computer Applications in the Biosciences"

2. "Computers and Biomedical Research"

3. "Computers in Nursing"

4. "MD-Computing"

5. "Medical Informatics".

"JAMIA" was not in existence when this work was done, having published its first issue in January 1994. Also included were two major conference proceedings for the medical informatics field, SCAMC and MEDINFO.

From a total population of 1676 articles, 271 randomly selected abstracts were reviewed, a 30 percent sample. The entire sample was reviewed by two members of the project team to determine the key words used in the abstracts. As a further check, two senior members of the project team reviewed 30 percent of the sample again. A weight of one was given to each term selected by the first team and a weight of two to each term selected by the second team because the members of the second team have more expertise in determining which terms are significant. These weights were assigned so that, once the entire raw vocabulary had been assembled, decisions could be made on what terms to eliminate based on the number of occurrences of the term and the number of times it was selected by each reviewer.

The key terms from each abstract were then collected and entered into a database and the editing process began. The initial editing consisted of eliminating singular vs plural, lexical variants, and useless terms. The decision was made to prefer the plural over the singular form of terms with some exceptions for singular words that are considered a preferred over their plural form; for example, classification was preferred over classifications and terminology over terminologies. In determining the preferred term for lexical variants the noun was preferred over the verb form and either of these was preferred over the adjective form again with some exceptions for preferred usage. The literature and the weights that had been assigned throughout the development of the raw vocabulary assisted us in determining the preferred variant of a term. A number of terms were eliminated because they were useless terms; for example, megabytes or house officer care or interactive nature. These terms carried weights of one or two which meant they had only been chosen from one article by one or both of the members of team one and upon examination by the senior team members were determined to be meaningless terms for the vocabulary. Also, the term for an action was preferred to the term for the person performing the action; for example, education was preferred over educators and development over developers. We also eliminated any hyphenated terms in favor of the non-hyphenated terms as in computer-assisted instruction or CD-ROM. Phrases with extra unnecessary terms were eliminated in favor of the shortest version applicable. Examples of this include the exclusion of automatic indexing method and automatic indexing program in preference to automatic indexing and the exclusion of diagnostic decision support system and diagnostic decision support program in preference to diagnostic decision support. From an original 5759 terms these edits reduced the terms to 5462. Weights were added together when a term was combined with another to produce a total reviewer weight.

The third step consisted of a review of glossaries from experts in the field to see what terms they used in their texts. Blois[10], Covvey[11], Shortliffe[12] and Barnett[13] were used as experts for this section. The glossaries were reviewed to ascertain which terms were already included in the raw vocabulary. Terms not already included were added to the raw vocabulary. A weight of two was assigned to terms in the Barnett and Shortliffe glossaries because they are more recent publications. A weight of one was assigned to the Blois and Covvey terms. Therefore, a new term appearing in all four glossaries would carry a weight of six.

Finally, the terms identified in each of the first three steps were compared to one another to see which ones were present in which sources. For those terms that existed in both the literature and the thesauri, the weights for each source were included in the database. There were some terms found only in the thesauri. If the thesauri weight of these terms was four or above, they were included in the raw vocabulary list. The terms found only in the expert glossaries were added to the raw vocabulary as bolded entries so that they would stand out when terms were reviewed. They were reviewed for what sources they belonged to as well as their weight. This insured that they would not be eliminated simply because they had a low weight.

RESULTS

When all the terms had been identified, the process of refining the raw vocabulary began. The aim was to reduce the original 5462 terms to a manageable size which Batty defines as "as small as possible but includes everything with a good size being approximately 2000 terms."[14] First, we closely examined the terms with a literature weight of 1 or 2 because that weight meant the term was selected by only one or two reviewers and from only one article. There were 3368 terms in this category. We eliminated 1738 terms, or 52%, of them in this step which left 3724 terms in the vocabulary.

The next step in the refinement process was to create a group of categories and begin to create the hierarchical structure for the classification. The main concepts then were:

1. Business
2. Computer and Data Processing
3. Education
4. Engineering
5. Language/Library
6. Legal
7. Mathematics
8. Nursing
9. Dentistry
10. Veterinary Medicine

All terms were sorted into one of these categories. Terms that did not fit were placed in a separate category for further review. Many of the latter terms were eliminated from the vocabulary as unimportant. Examples of terms eliminated include crime, shared data, and system environment. The major categories were refined in the second step of this process to include:

1. Business
2. Computer and Data Processing with mathematics and engineering as sub-categories
3. Education
4. Health Care Informatics which includes the health category
5. Language/Library

The categories of legal, nursing, dentistry and veterinary medicine were absorbed into the other categories.

Once the terms were divided into categories, we began to establish a hierarchy of terms within each category. At this point, preferred terms were chosen for concepts which had several synonyms. Part of this final refinement was to decide how acronyms and geographic terms were to be handled. Not all acronyms were retained in the vocabulary. For those we did retain, it was decided to put the acronym in parentheses behind its spelled out version wherever it appeared in the thesaurus. Examples of this would be Computer Assisted Instruction (CAI) or Integrated Advanced Information Management Systems (IAIMS). For geographic names users of the thesaurus will be referred to the Z Tree of MeSH which contains a comprehensive listing of geographical place names.[15] Users will also be referred to the MeSH Subject Headings for specific medical terms as the purpose of this project was not to recreate the entire medical vocabulary but only to include terms that specifically addressed informatics and its related fields. The decision to handle acronyms, geographic names, and specific medical terms in this manner eliminated 1378 terms which gave us a total of 2315 terms remaining in the vocabulary.

When completed, our medical informatics thesaurus will contain a manageable number of terms, arranged within a hierarchical structure within the final five categories. The thesaurus will then be tested against the 1994 SCAMC titles to see if it contains the necessary terms to cover the concepts in the document. We also plan to test the document by sending it to experts in the field for their review, use and comments. Finally, we will publish the thesaurus in both a print and an electronic version.

DISCUSSION

The first formal work in the terminology of medical informatics was at NLM. They began to cover medical informatics terminology in a minimal way in their first L Tree developed in 1960. Major revisions to this L Tree in 1963, 1965, 1966, 1975, and 1987 have added new terms, eliminated terms and changed the hierarchical relationship of terms.

Rada and others[16] helped in this revision process in 1986 when they developed a medical informatics thesaurus. "It consisted of terms developed by an automatic merging of the thesaurus used by the "Association of Computing Machinery" and the Information Sciences component of the "Medical Subject Headings" from the National Library of Medicine (NLM). The terminology was then pruned by eliminating terms not related to those in the MEDINFO keyword list or not in the medical informatics literature." The terminology from this thesaurus was incorporated into NLM's 1987 version of the L Tree.

Currently Rada is working with the Committee for European Normalization (CEN) under the International Medical Informatics Association (IMIA). This work parallels our project. [17],[18] They have produced a new 200 word thesaurus.

Rada's terminology still leaves a wide area of informatics unrepresented since it is focused on the creation of a framework for standards development. The Rada et. al. work and this project also differ in methodology of thesaurus construction and size of the vocabulary.

Rada uses a computerized technique to extract possible terms. This method achieves literary warrant but makes little formal effort to ensure user warrant. He has published the list and requested comments from members of the informatics community. Our methods of construction were more extensive and varied and achieve both user and literary warrant. As shown, our approach uses literature review, thesauri review, review of expert glossaries, and comparisons of all sources for similarity to achieve both literary and user warrant

for our terminology.

The small size of Rada's vocabulary (200) necessarily would force users to use broad terms for some concepts. With a larger thesaurus of 2000 or more terms the user is more likely to be able to achieve an acceptable level of specificity.

CONCLUSION

Our research first identified a need for a new medical informatics thesaurus. Then, using existing thesauri, medical informatics literature, and the terminology of experts in the field to identify appropriate concepts and terms, we are creating a new thesaurus to cover the discipline. Continued refinement will be accomplished by testing the thesaurus against the 1994 SCAMC titles and by asking experts in the field to use and review the thesaurus and give us feedback. The thesaurus will be published in both a print and an electronic format so that it is readily available for use by everyone in the field. We plan to support continuous revision of the thesaurus to keep it current with the ever expanding field of medical informatics.

References

1. Li Z, Ogg N, Sievert M, and Mitchell J. "On the Growth and Trimming of the L Trees of MeSH." Symposium on Computer Applications in Medical Care. 1993: 892.

2. Lancaster FW. Vocabulary Control for Information Retrieval. lst ed. Arlington, VA: Information Resources Press, 1972: 27.

3. Li Z, et al, 892.

4. Milstead JL. ed. Engineering Information Thesaurus. 1st ed. Hoboken, NJ: Engineering Information Inc, 1992.

5. Institution of Electrical Engineers. INSPEC Thesaurus. Old Woking, Surrey, Eng: Unwin Brothers Limited, 1991.

6. Houston JE. ed. Thesaurus of ERIC Descriptors. 12th ed. Phoenix, AZ: ORYX Press, 1990.

7. Walker A. Jr. ed. Thesaurus of Psychological Index Terms. 6th ed. Arlington, VA: American Psychological Association, 1991.

8. National Library of Medicine. Medical Subject Headings. Annotated Alphabetic List. Bethesda, MD: Medical Subject Headings Section, Library Operations, National Library of Medicine, 1992.

9. Lotus Development Corporation and Ziff Communications Company. Computer Select® Glossary of Terms. 5th ed. New York, NY: Ziff Communications Company, 1992.

10. Blois MS. Information and Medicine: The Nature of Medical Descriptions. Berkeley: University of California Press, 1984: 298 p.

11. Covvey HD, McAllister NH. Computers in the Practice of Medicine. Vol. I-Introduction to Computing Concepts. Reading, MA: Addison-Wesley, 1980: 266 p.

12. Shortliffe EH, Perreault LE, eds. Medical Informatics: Computer Applications in Health Care. Reading, MA: Addison-Wesley, 1990: 715 p.

13. Barnett GO. "Core Topics in Medical Informatics." Personal Correspondence, 1992.

14. Batty CD. Thesaurus Construction Workshop. 55th Annual Meeting of the American Society for Information Science (ASIS). Oct. 22, 1992; Pittsburgh, PA.

15. National Library of Medicine. Medical Subject Headings. Tree Structures. Bethesda, MD: Medical Subject Headings Section, Library Operations, National Library of Medicine, 1960-1992.

16. Rada R, Calhoun F, Mili E, Singer SJ, Blum B, and Orthnerr H. "A medical informatics thesaurus." MEDINFO '86. Washington, DC. Amsterdam: North-Holland, 1986: 1164-1172.

17. Rada R, Ghaoui C, Russell J, and Taylor M. "Approaches to the construction of a medical informatics glossary and thesaurus." Medical Informatics. 18(1). 1993: 69-78.

18. Rada R. "Vocabulary." SigBio. 14(1). Jan. 1994: 5-16.