MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

The MeSH Translation Maintenance System: Structure, Interface Design, and Implementation

Stuart J. Nelson[a], Michael Schopen[b], Allan G. Savage[a], Jacque-Lynne Schulman[a], Natalie Arluk[a]

[a]*National Library of Medicine, Bethesda, MD, USA*
[b]*Deutsches Institut für Medizinishe Dokumentation und Information, Köln, Deutschland*

Abstract

The National Library of Medicine (NLM) produces annual editions of the Medical Subject Headings (MeSH®). Translations of MeSH are often done to make the vocabulary useful for non-English users. However, MeSH translators have encountered difficulties with entry vocabulary as they maintain and update their translation. Tracking MeSH changes and updating their translations in a reasonable time frame is cumbersome. NLM has developed and implemented a concept-centered vocabulary maintenance system for MeSH. This system has been extended to create an interlingual database of translations, the MeSH Translation Maintenance System (MTMS). This database allows continual updating of the translations, as well as facilitating tracking of the changes within MeSH from one year to another. The MTMS interface uses a Web-based design with multiple colors and fonts to indicate concepts needing translation or review. Concepts for which there is no exact English equivalent can be added. The system software encourages compliance with the Unicode standard in order to ensure that character sets with native alphabets and full orthography are used consistently.

Keywords:

MEDLINE; Translations; Subject Headings; Unified Medical Language System; Databases; User-Computer Interface

Introduction

Background

The National Library of Medicine's (NLM) MEDLINE database includes over 13 million literature citations of articles written in 41 languages [1]. Each article is indexed with Medical Subject Headings by an individual who, after scanning the article in its original language, assigns the descriptors to indicate what the article is about.

New editions of the Medical Subject Headings [2] are produced annually. Editing may alter a heading to a new, more modern usage, revise the organization of the headings, add new headings to cover new topics, add entry vocabulary, or otherwise modify the previous release. After each new edition is produced, the MEDLINE data is made consistent with the new headings algorithmically, replacing old headings, occasionally simply adding a new heading, and rarely deleting an old heading for which there is no suitable replacement [3].

For many years, national medical information centers outside the United States have produced translations of MeSH to make the vocabulary useful for non-English users. These translations varied in frequency of appearance. Some translations were issued annually and others irregularly. Translations of MeSH have been made into Arabic, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Portuguese, Russian, Romanian, Thai, Turkish, Slovene, Slovak, Spanish, Swedish. Other translations, that we are not aware of, may have been done as well.

Concept Structure of MeSH

In the year 2000, MeSH changed from a database of descriptors and terms, to one consisting of descriptors, concepts, and terms [4]. A descriptor is now viewed as a class of concepts, and a concept as a class of synonymous terms within a descriptor class. By using the concept as a key object in the new structure, appropriate non-synonymous relationships could be represented separately, and differences between usages and meanings clarified and disambiguated. The descriptor class consists of one or more concepts closely related to each other in meaning, or of non-synonymous concepts best lumped together in one class for the purposes of indexing, retrieval, and organization of the literature. Putting these non-synonymous concepts together into one descriptor class does not alter the traditional function of entry vocabulary, that of pointing the user (whether individual or system) to the appropriate main heading (descriptor). Rather, it points out explicitly that this was a choice for the intended purpose of the vocabulary, rather than a confusion about the meaning of a term.

For example, under the old term-based system, the descriptor CYTOPLASMIC GRANULES had a non-synonymous (narrower) entry term, Secretory Granules (Figure 1). After establishment of the new concept-oriented system, it happened that a new descriptor SECRETORY VESICLES was created. While CYTOPLASMIC GRANULES was retained as a MeSH main heading, Secretory Granules was moved to this new heading as a (related) subordinate concept, better represented by the new descriptor. Other non-synonymous subordinate concepts were also created under this new descriptor class. The multiple, non-synonymous concepts represent slight differences in meaning,

yet they are all grouped together under SECRETORY VESICLES for purposes of retrieval (Figure 2).

Old Main Heading: CYTOPLASMIC GRANULES
Old Entry Term:Secretory Granules

Figure 1 - Heading before Changes

Descriptor: CYTOPLASMIC GRANULES
Descriptor: SECRETORY VESICLES
Preferred Concept: Secretory VesiclesSubordinate Concept: Secretory Granules
Subordinate Concept: Condensing Vacuoles
Subordinate Concept: Zymogen Granules
Subordinate Concept: Dense Core Vesicles

Figure 2 - Headings after Changes

The change in data structure allows a greater degree of organization. Each descriptor class has a preferred concept. The term that names the preferred concept (the preferred term of the preferred concept) provides the name for the descriptor. Each of the subordinate concepts also has a preferred term, as well as a labeled (broader, narrower, related) relationship to the preferred concept. Terms meaning the same (naming the same concept) are grouped together in the concept record.

Translation Database

Translations serve not only as a way for national centers to organize information not covered in NLM databases, but also serve an important function for MEDLINE users not facile in English. It is far easier to search for information in a language with which one is very familiar. When articles that are of sufficient potential interest to warrant closer inspection are found, the effort necessary to read the article can be made. The citation maintenance practices used in MEDLINE make the updating of the translations highly desirable.

For these and other reasons, it was felt there was need to extend the MeSH maintenance system to encompass an interlingual database of translations. The MeSH Translation Maintenance System (MTMS) allows continual updating of the translations, as well as facilitating tracking of the changes within MeSH from one year to another. In this way translations can be made on new headings as they are created rather than waiting until after they are published once each year, greatly facilitating currency. Given the new MeSH structure, it was relatively easy to conceive of a method for supporting the work of translators. Translated terms can be included in the MeSH maintenance environment as an extension of the current MeSH database. Translators can use the MTMS to manage their translations.

Translated terms are provided as synonyms to existing concepts. For non-synonymous entry terms that are not present in the English version, but useful in the language of the translation, the translator creates a new concept. The new concept would, of course, belong to a descriptor class, that of the main heading for which it was an appropriate entry term. In this case of a concept class for which there is no English synonym, a definition of the concept is required, so that translators using other languages can have the ability to include their terms in that concept class.

Materials And Methods

Design Of The Interface

In order to avoid the difficulties of trying to maintain clients an multiple disparate platforms scattered across the globe, the interface was designed to be Web-based. A variety of security measures limit participation to authorized individuals. ColdFusion templates, running on the Web server, enable the transmission of the submitted information to the database server.

Privileges for translators are limited to insertion of terms in their own language, and to creation of new subordinate concepts. In the case of creating a new subordinate concept, the required submission of a definition (in English) of the new concept supports both the translation of that term into other non-English languages, and enables proper maintenance when that descriptor class is edited by the MeSH staff.

While the translator has the ability to browse MeSH descriptors, the translation interface has been designed for direct editing of concepts and terms only. There are two different ways that the user can access the concepts and terms from the interface: (1) by navigating the MeSH Tree Structures to descriptor names or (2) by searching on an individual term. For each method, there are two different language modes available: an English version, and a translated version that appears in each users own language.

The interface uses color, boldface, and italic fonts in the display to convey the current status of the various descriptors, concepts, and terms. In this way the user can quickly determine at a glance which MeSH terms are new, which still need to be translated, which have been determined to be untranslatable, and which translated terms are waiting supervisor review and final approval.

A special module of the interface was designed for supervisors of each translation. This module allows the supervisor to review and to authorize work done by their group.

Managing The Workflow

For each language incorporated into the MeSH maintenance environment, it is anticipated there will be a team of translators and a supervisor. The supervisor will coordinate, review, and authorize the work of that group of translators for that language. Once the supervisor authorizes the work, a member of the MeSH Section at the National Library of Medicine will conduct a final review and quality control of all changes before they are approved to become an official part of MeSH.

To institute the translation database requires the agreement and cooperation of the translators. Once that is obtained, any prior work will be loaded into the MTMS. Some of the current translations have been incorporated in the UMLS Metathesaurus. Those translations, in concept structure, can easily be incorporated into the MTMS. Translations that have not been previously included in the UMLS Metathesaurus will be dealt with on an individual basis. After the translations are loaded, translators

will then be able to review areas in which the mapping from one term to another might be problematic, and to find the descriptors in MeSH for which there was no translated term. The display of translated terms in the concept structure allows careful review to be sure that the finer shades of meaning are fully represented.

Character Set Issues

The character set used depends on the operating system and the coding scheme it uses for the language. Knowing the coding scheme, it is often possible to find a set of fonts (or glyphs) to make the character appear as it should in the specific written language. However, the coding schemes are not unique, and far from universal. Unless the scheme is understood properly, sorting and presenting material in an orthographic manner becomes quite difficult.

The best long-term solution to the character set problem is one that correctly represents languages with their native alphabets and full orthography. Unicode [5] appears to be one means of achieving that goal. It provides a unique number for every character, no matter what the platform, program, or language. The MeSH database has been converted to Oracle version 8I, a database management system that supports the use of Unicode. Java, which supports the template and the MeSH client used at the NLM, is fully Unicode compliant. The MTMS will encourage translators to submit their terms in Unicode.

When the source file of terms in another language is loaded into the MTMS, Oracle 8I converts the coding (Unicode or otherwise) for each character to UTF-8 (Unicode Transformation Format-8), which is how they are stored in memory. The Web server, in conjunction with the MTMS application and the IE Browser, is also configured to UTF-8 encoding. In this way, a consistent character set with native alphabets and full orthography is used and conforms to a universal standard.

Discussion

Part of the UMLS® project [6], the UMLS Metathesaurus® is a large database of naming information encompassing terms and concepts from more than 50 biomedical vocabularies and classifications, including MeSH. The Metathesaurus is updated on a regular basis. A number of translations of MeSH are included in the UMLS Metathesaurus. The translations into German, provided by DIMDI, French (INSERM), Portuguese (BIREME), Spanish (BIREME), Russian (Central Medical Library, Moscow), Italian (Istituto Superiore di Sanita), and Finnish (the Finnish Medical Society) are included in the 2003 Metathesaurus. Without specific links between the translated terms and English terms in MeSH, maintaining appropriate representation in the Metathesaurus requires considerable effort on the part of the translators or bilingual editors. While an original translation of all of MeSH might link a term with the appropriate subject heading, with modifications of MeSH the linkage of that translated term might no longer be appropriate. Non-synonymous entry terms account for most of the problems thus generated. Entry terms for which there are no English synonyms are also problematic.

Each year, during the summer, the final version of the new MeSH descriptors for the following year is completed. Files of this new version are prepared for inclusion in the UMLS Metathesaurus, for citation, and for distribution. The present practice is that the translators of MeSH who contribute their translations to the UMLS then receive the lists of new and changed descriptors. They are then obligated to provide, within a few months, their updated translation. With concurrent translations enabled by the MTMS, the work would be spread over the year. Then, at some time shortly after the completion of the new version of MeSH, files can be prepared of each translation for transmittal back to the national centers.

We anticipate that the MTMS will allow for much easier inclusion of MeSH translations into the UMLS Metathesaurus. A uniform format, a common character set, and the linking of all terms to MeSH Concept Unique Identifiers, along with the previously established links to the Descriptor Unique Identifiers will facilitate the preparation of each vocabulary for insertion into the Metathesaurus.

The existence of the MTMS should reduce the barriers to developing translations and keeping the translations current. National centers without significant resources would be relieved of the problem of providing software support for translators. Further tasks in making MEDLINE more accessible, such as defining a search interface in the native language, can then be approached.

Conclusion

Translations of MeSH provide an important mechanism for individuals not familiar with English to access MEDLINE. The MTMS supports translators, enabling correct mappings from one language to another to be maintained and to be current with MeSH. The Web-based interface, closely managed maintenance environment, and adherence to modern standards, all provide a robust platform for an interlingual database of translations.

References

[1] MEDLINE [electronic resource] / National Library of Medicine. [Bethesda, MD : The Library], 1966-

[2] Medical subject headings. U.S. Dept. of Health and Human Services, Public Health Service, National Institutes of Health, National Library of Medicine ; [Washington, D.C. : Supt. of Docs., U.S. G.P.O., distributor].

[3] Humphrey SM. File maintenance of MeSH headings in MEDLINE. *J Am Soc Inf Sci.* 1984 Jan;35(1):34-44.

[4] Johnston, Douglas; Nelson, Stuart J.; Schulman, Jacque-Lynne; Savage, Allan G.; Powell, Tammy P. Redefining a Thesaurus: Term-Centric No More. Poster presentation at: *AMIA 1998 Annual Symp.*; 1998 Nov 10; Orlando FL.

[5] Unicode Home Page [Internet]. Mountain View (CA): Unicode, Inc; c1991-2001 [cited 2003 Sept 15]. Available from: http://www.unicode.org/.

[6] 6. Humphreys BL, Lindberg DA. The UMLS project: making the conceptual connection between users and the information they need. *Bull Med Libr Assoc.* 1993 Apr;81(2):170-7.