A Model International Cooperative Project – The MeSH Translation Maintenance System

Jacque-Lynne Schulman and Stuart J. Nelson, MD
National Library of Medicine, Bethesda, MD, USA

Abstract

Medical Subject Headings (MeSH) is created and maintained by the US National Library of Medicine (NLM). A major biomedical vocabulary, it is used world-wide by information scientists working in many languages. Over the last decade, an increasing number of translations of the MeSH have been prepared by centers outside the United States. Translations are used to index literature not indexed by the NLM as well as to provide native language interfaces to MEDLINE. MeSH contains more than 23,000 descriptors and ten of thousands of entry vocabulary. It is produced annually, with numerous additions, changes, and deletions. Tracking these changes and maintaining consistency with the current version of MeSH is often a challenge for a translator.

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) allows continual updating of the translations, as well as facilitating tracking of the changes within MeSH from one year to another. This system has been in use at several institutions for the last year. This paper reviews the requirements set for the translation system, describes the design of the system, and reports on the first uses.

Technical support and computer resources used to manage and organize the translations data and produce the data in the format required by the translation partners are provided by the NLM with no expense to the translation partner. The translation partners provide their language expertise and local scientific knowledge in the intellectual tasks that form the content of the MTMS. Partners assume the responsibility of implementing the translations in their own systems. This cooperative approach provides access to current work on MeSH and reduces the costs of maintaining translation systems for the translators.

Keywords:
MEDLINE; Translations; Subject Headings; Unified Medical Language System; Databases

Introduction

Background

The National Library of Medicine's (NLM) MEDLINE database includes over 14 million literature citations. The citations represent articles written in 41 languages [1]. Each article is indexed with Medical Subject Headings by an indexer who, after scanning the article in its original language, assigns the descriptors to indicate what the article is about.

Deleted: million literature

New editions of the Medical Subject Headings [2] are produced annually. The main thrust of a new edition is to add new headings to cover new topics or to add entry vocabulary. Revisions, either altering a heading to a new, more modern usage, or to revise the organization of the headings, are also a feature. After each new edition is produced, the MEDLINE data is made consistent with the new version of MeSH [3]. Each citation is automatically checked for necessary changes to bring its descriptors up to date with all vocabulary changes.

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. 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.

The increasing dominance of English language publications has been noted [4]. The predominant language of publication in PubMed has been English and the dominance is increasing. For researchers and health care practitioners whose first language is not English, the ability formulate questions and access data in their preferred language, is ever more important. Support of a robust and current array of translations of MeSH is one more way in which there is equity of access, regardless of language, location, or nationality. When articles that are of sufficient potential interest to warrant closer inspection are found, the effort necessary to read the article or obtain a translation can be made.

Deleted: -
Deleted: The following data ??
Deleted: confirms that
Deleted: t
Deleted: that

The translations have varied in frequency of appearance. Some translations were issued annually and others irregularly. Translations of MeSH have been made into Arabic, Chinese, Dutch, Finnish, French, German, Greek, Italian, Japanese, Polish, Portuguese, Russian, Romanian, Thai, Turkish, Slovene, Slovak, Spanish, and Swedish. Other translations, that we are not aware of, may have been done as well. Interest in translations appears high, with considerable discussions, and a number of projects started, over the past decade. The citation maintenance practices used in MEDLINE make the updating of the translations highly desirable. As the citations are updated annually in line with annual changes in the MeSH vocabulary, to be most effective, a translation should also be updated annually to include all changes in MeSH.

Concept Structure of MeSH

In the year 2000, MeSH changed its maintenance environment. In changing from a database of descriptors and terms, to one consisting of descriptors, concepts, and terms [5], it became possible to support concept-based translations. 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 are represented separately, and differences between usages and meanings are 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 choice is intended to serve the purpose of the vocabulary, instead of being confusion about the meaning of a term.

Deleted: of being

For example, under the old term-based system, the descriptor CYTOPLASMIC GRANULES had a non-synonymous (or 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 the 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.

The Translation Database
With the change of the data structure to be concept-oriented, rather than term-oriented, a translation database became feasible. Translators can use the MTMS to manage their translations. Instead of tracking terms as they move within MeSH through the years, possibly with evolving meanings, a translation database allows the translator to attach their term to the precise meaning they wish to express. Movement of concepts within the larger structure of descriptors becomes automatic for the translators.

The MeSH Translation Maintenance System (MTMS) also allows continual updating of the translations. 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. In the past, there was no way to know what changes were in progress between the annual distribution of the MeSH vocabulary. The MTMS system makes the ongoing vocabulary work visible to the translators. Given the rate of change of MeSH, with a large number of additions made annually, more time is now available to complete the translation. For example, for the 2006 version of MeSH, there are over 900 new descriptors. The time available for creating the translation is the entire production year, rather than a 6 week window period between the completion of the new version of MeSH and its institution within the NLM systems.

Experience with the MTMS
Initialization
The MTMS was opened for use in 2003. After an initial test period involving the German translators, other users were welcomed. At the present time, French, German, Czech, Finnish, and Russian translations are supported in the MTMS. Japanese, Slovenian, Vietnamese, Polish, Chinese, and Korean are at various states of beginning efforts in the MTMS.

For the most part, the initialization of the MTMS for a translator is fairly straightforward. A previously existing translation is loaded into the MTMS. In general, it is easier with a term-by-term translation or a concept-based translation. Translations loosely based on descriptors are more problematic, and require review by translators after insertion into the MTMS. Tree Categories, which are not formally part of MeSH, are translated in the MTMS by the translators as they begin the use of the MTMS.

Character Set Issues
The participants in the MTMS cooperative venture have used various system and language coding arrangements in the past. One of the outcomes of the MTMS effort is a uniform approach to character

Deleted: ¶
Formatted: Font: Not Bold
Deleted: ¶
Deleted: is loaded
Deleted: is easier

representation. Previously, some translations were upper case (capital case) only and some also lacked representation of necessary diacriticals. There were also locally developed coding systems.
In the absence of a standard and universal approach, exchange of translations data was difficult and incorporation into the UMLS even more difficult.

The best long term solution to the character set problem is one that correctly represents languages with their native character sets and full orthography. Unicode 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 runs in Oracle version 8I, a database management system that supports the use of Unicode [6]. Java, which supports the servlet and the MeSH client used at the NLM, is fully Unicode compliant. Using the MTMS, the resulting translations are created in a common and internally used standard coding system.

When the source file of terms in another language is loaded into the MTMS, the database system, 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, and full orthography including all diacritics, is used and conforms to a universal standard.

Most translations have focused on the descriptors, usually providing names for the preferred concept, but adding other terms and concepts as desirable. One group found it desirable to translate scope notes. Although that had not been part of the plan for the original system, it was found to be an easy addition to the system.

Timing Issues
At any one time, there are two versions of MeSH outstanding, on which a translator may wish to work. There is the current version, usually the one in use in the PubMed database, and the new version, which reflects the active work of the MeSH staff. Annually, there is a period of MeSH production known as cutover, during which the new version is authorized and approved for use. It then becomes the current version. Because few translators have completed their translation at the same time as MeSH cutover, it is necessary to cutover the translations at a later date, before the new version of MeSH comes into use, but after the MeSH cutover has taken place.

Another interesting aspect of timing is the lag in communications. E-mail from Europe arrives early in the day. By the time it has been answered, it may be the next day before it is read there. And, as any collaborative group well knows, it is hard to do it all by written communication. Telephone conversations are also affected by time differences and technical issues may be actually easier to communicate in writing.

Official Versions
It is often the case that there are differences between countries and continents in the medical or scientific meanings of terms in widely spoken languages. Political considerations about the official nature and the ability to use a translation into a certain language may also occur. The MTMS takes no position on how official any version of the translation may be. It will support multiple translations into the same language, by identifying the institution or source of the translation.

Returning Files
Almost every group doing translations has asked for some slight differences in having their work returned to them. Some would like monthly updates, others are content with an annual update. Overall, the format most commonly used to return the translation has been an xml version, based on the MeSH

xml DTD. All can see their progress online as their work proceeds in the MTMS. Additionally, translators are allowed to view any work done by other translators on the descriptor on which they are working.

Futures

Several of the translating groups have used their translations to make both their own material as well as MEDLINE available through a single interface. The groups index a number of journals in their own language. The search interface then allows the user to look at locally indexed material as well as the MEDLINE material.

Of interest is that several of the translators find it desirable to add new descriptors to cover concepts of importance for local indexing, but which are not necessary for MEDLINE. An example of this is the expanded coverage of place names in Korean. Most of these are Korean cities and provinces. While important in a Korean context, in the worldwide context these names are very finely granular. Another example is the potential use of MeSH to index information about Chinese Herbal Medicine, which has a thesaurus based on MeSH [7]. There are a considerable number of terms for herbal medicines which are not presently in MeSH, as well as diagnostic categories and other methods of describing patients. At some point in the future it may very well be desirable to provide an enhancement to the MTMS in which building an extension could be supported, even while the links between the MeSH base and the extension are kept current.

The Model

Deleted: ¶

We believe that the MTMS provides a good working model for supporting a multilingual collaborative translation effort. While providing the base MeSH as part of its ongoing mandate, the NLM can also support those from outside the USA who wish to avail themselves of the important resources indexed with MeSH. The NLM can make use of its resources in an environment which allows others to see the work in progress, and apply their expertise and language skills.

Conclusion

Deleted: ¶

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.

The ability to support supplementing of MeSH with terminology of local importance such as place names, medical conditions with regional focus, and the like, appears to be a likely target of future efforts.

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. Loria A, Arroyo P. Language and country preponderance trends in MEDLINE and its causes. Med Libr Assoc. 2005 Jul;93(3):381-385.

5. 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.

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

7. Wu, Lancheng. Zhongguo Zhong yi yao xue zhu ti ci biao [Chinese traditional medicine and materia medica subject headings]. Beijing : Zhong yi gu ji chu ban she, 1996.

Formatted: Font: (Default) Times New Roman, 12 pt

Deleted: TCM thesaurus¶