Integration of a Standard Gastrointestinal Endoscopy Terminology in the UMLS Metathesaurus

Michele Tringali, M.D.[a], William T. Hole, M.D.[b], Suresh Srinivasan, M.S.[b]

[a]Gastrointestinal Endoscopy Unit, Ospedale Regionale della Valle d'Aosta, Italy[1]
[b]National Library of Medicine (NLM), Bethesda, MD, USA

MST©, a standard terminology for gastrointestinal endoscopy reporting, was integrated in the January 2002 UMLS Metathesaurus in order to ease the practical interoperability of clinical data repositories in gastroenterology. The integration required full specification of names, resolution of discrepancies between English, French and Italian versions of MST, appropriate categorization with UMLS Semantic Types and MST-level Class attributes, assignment of explicit intra-table (and some useful inter-table) relationships mainly at concept level but also at the source level in order to retain and fully represent the original explicit and implicit MST organization. Main results, problems encountered and future plans are discussed.

INTRODUCTION

The goal of the NLM's UMLS project is to integrate information from various sources, including clinical records, so as to improve access to biomedical information both for clinicians and patients. This could be achieved only if patient-description vocabularies able to represent data in the same detail as used in progress notes are developed and integrated into the Metathesaurus[1].

The 2001 edition of the UMLS[2] does not include a terminology adequate to the gastrointestinal endoscopy (GIE) specialty. The Minimal Standard Terminology (MST©)[3] is an authoritative controlled list of preferred terms to be used for description of "Reasons for performing the endoscopy", "Findings", "Endoscopic Diagnosis" and other details of examination in GIE reports. The terminology has been prepared jointly by the European Society for Gastrointestinal Endoscopy (ESGE) and American Society for Gastrointestinal Endoscopy (ASGE) and has been validated in prospective tests[4]. MST is available in the public domain in ten national languages: Czech, English, French, German, Hungarian, Italian, Portuguese, Russian, Spanish, Turkish.

MST is more a controlled list of preferred terms than a full fledged terminology. The major strengths of the MST are its completeness and flexibility: it aims to cover all the sections of the endoscopic record that have clinical importance. Its area of weakness are the lack of allowable synonyms, the absence of suitable mapping to non-proprietary reference terminologies, and insufficient modular and concept-based structure that limits its practical usability in applications and even in research[5].

Our aim was to achieve a mapping of the MST terms into the UMLS Metathesaurus in order to provide a new (but MST-compatible) machine-readable terminological tool that could be used to appropriately link GIE related data to many different types of biomedical information systems.

METHODS

1) Data gathering and analysis

The tables of the MST were extracted from the official versions for the English and Italian languages. The French translation of the MST, not yet published, was received from the General Secretary of ESGE. For each national language, the 24 original MST tables were saved as RTF files on a MS Windows 2000© PC.

MST encompasses 1713 symbols in its titles and tables sections: 122 Reasons, 8 Endoscopic Procedures, 1030 Findings, 7 Complications, 166 Additional Procedures, 235 Diagnosis, 93 Sites (anatomical), and 52 Details of the examination.

The structure of MST relies on 24 tables (Figure 1), nearly all of which are organized in a five-column structure: *Heading*, or general class; *Term, Attribute, Attribute Value,* and *Site*. Embedded in this structure is the informational model of the GIE record. For example, in the "Findings" section five Headings, or categories of Terms, are provided: *Lumen*, *Content*, *Flat*, *Protruding*, or *Excavating lesions.*

This organization leads to an almost complete absence of fully specified terms (terms that do not require other, or contextual, information to be fully understood by an agent e.g. a human reader or an "intelligent" computer program). As an example, in order to express the concept of "superficial gastric ulcer" using the terms of MST as reported in the following extract from the allowable *Attributes* and

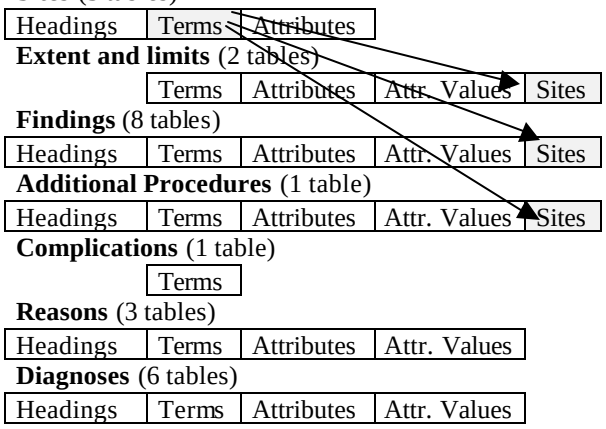
[1] MT is now at Knowledge Centre (Unità Conoscenza e Ricerca), Azienda Ospedaliera S.M. Misericordia, Udine, Italy

Values in Table 7 of MST:

Headings	*Terms*	*Attributes*	*Attr. Values*	*Sites*
Excavated lesions	Ulcer	Shape	Superficial	Site (s)

an agent must combine an *Attribute Value* (Superficial) with a *Term* (Ulcer) and then use a method to refer the new string to the title of the table, e.g.: "superficial ulcer (of the) stomach".

Fig. 1: Structure of MST and a specified relationship



Explicit relationships are also lacking in the original MST tables. Many relationships are implied in the structure of the terminology, though these are not consistently applied.

A number of linguistic and structural problems (and some errors, for example, attribute-values with different meaning in different sources and un-translated English terms in the French translation) were found during the process of alignment of the three language versions (Table I). They were addressed with a detailed analysis, applying linguistic and domain knowledge. These discrepancies, along with actions taken while preparing the inversion process, are listed in a file that will be available on NLM website.

2) Inversion of the source

Source inversion is a complex process aimed to preparing the objects of a terminology (files, tables, hierarchies, annotations) in a format suitable to the insertion into the Metathesaurus.

The inversion process for MST was not straightforward and required several iterations of human review. The work started with the English files. Terms from the Italian and French files were added later as synonyms of the English terms.

Eight of the 24 tables (the "Findings" section of the MST) were converted to plain text files using MS Word© and sent to Apelon, Inc. (http://www.apelon.com) for automated processing. The remaining 16 tables were manually worked by the primary author, since their structure was different.

Table I: A sample of discrepancies between English, French and Italian versions of MST.

Difference in granularity in 1/3 or 2/3 languages (L)
"Other" terms in 1/3 or 2/3 L only
Inversion of couples (or more) of entries
Synonyms marked by / or parentheses in 1/3 - 2/3 L
"Specify" as Term instead of Attribute Value in 1/3 L
Lacking of an attribute or an excess attribute in 1/3 L
Attribute misplaced in attribute value column (or the opposite)
Lacking term-attribute-value triplet in 1/3 L
Attribute-values with different meaning in one source
Two attribute values condensed in one entry in 1/3 L
Two more specific attribute values in 1/3 L
Misplacing of a triplet (term-attribute-value)
Un-translated term from another L version

The goals of the process were:

- to assign a unique and fully-specified name ("atom name") to all the pertinent terms in the source
- to relate each fully-specified "atom" to its original term, using a mapping method that could assure easy maintenance of the source, and that could faithfully represent the content of the MST tables
- to assign appropriate Semantic Types (STYs) to terms, with STYs names and meaning derived from the UMLS Semantic Network (SN)
- to assign appropriate MST_Classes (source-level attributes) to each fully-specified name, in order to retain as much as possible the implicit semantic categorization of the original MST tables
- to make explicit the implicit relations among the terms in the original tables, assigning adequate source-level relationships to the atoms. Wherever possible, these relationships were to be drawn from the UMLS SN, preserving consistency with the SN definitions and usage notes
- to add a limited number of useful inter-table relationships, expressing the relations between Findings and their allowable Sites.

At the end of the inversion process, a MST-compatible but completely new informational structure, named MTHMST2001, was produced.

Full specification of names

An algorithm was developed by Apelon following specifications from the primary author, and it was applied to the (longer, most structured) "Findings" tables in order to extract a pseudo-fully specified name (an unsupervised fully specified term to be later manually revised for validity and clarity) for each

term. The algorithm was designed to concatenate the content of selected fields in appropriate order, and to add the adjectival form of the table title. The processed files were then manually edited with the goal to produce good natural language surface names for any pseudo-fully specified term. This manual revision required much effort, since no more than 1/3 of the unsupervised terms produced by the algorithm could be used as fully specified names without further editing.

Fully specified atoms (and relations) were manually produced for the shorter "non-findings" tables, checked for duplicates, and edited accordingly.

The full specification process aimed at producing names both expressive enough to mirror the actual use in records, and uniquely linked to the underlying unambiguous concepts. A singular form was chosen even when the MST had the plural one; British spellings were replaced by American equivalent. The first occurrence of "Normal" in each table (recurring two times at the beginning of each file) was substituted with the notation "Finding" preceded by the adjectival form of the table title, e.g. "Esophageal finding", "Gastric finding".

A high degree of redundancy was retained even for questionable terms, to respect faithfully the source hierarchies, e.g. the related terms: "Bleeding", "Yes", "No" in table 8 (terms for duodenal findings) were fully specified as: **Bleeding of Hemorrhagic Duodenal Mucosa, Presence of Bleeding of Hemorrhagic Duodenal Mucosa** and **Non Bleeding Hemorrhagic Duodenal Mucosa**. Semantic distance between the first and the second term is obviously really subtle, but they were left for consistency with the source organization.

From a lexical viewpoint, the method used to fully specify a name was geared to minimize the production of strings not useful for natural language processing[6]: embedded classification features, underspecification, bracketed expressions, inverted strings (such as strings with a comma followed by a space). For 134 strings, out of the total number of 1945, one of the unsuitable features could not have been avoided and this will be addressed in a future work involving an usability test.

Some examples of original MST names (not fully specified; in "quotes"), algorithmically derived names (pseudo-fully specified; in *italics*) and manually edited names (complete full specification; in **bold**) are reported in Table III.

For the French and Italian versions of MST, no algorithm was applied even to the "Findings" tables, since it was judged easier to apply language-specific linguistic knowledge to manually specify the names using, as a template, the English fully specified names listed in a simple MS Word table.

Mapping of fully-specified names to original unspecified strings

Titles of tables and each non-empty value in the original tables were considered to be MST strings (with the exception of the "Site(s)" repetitive notation in Findings, Details of the Procedure and Additional Procedures tables, that were converted to explicit relationships; see below). Each MST string was linked to its fully-specified MTHMST2001 atom using an attribute field, with information on its position in the tables stored as a concatenated text string, in the following format: *number of table & number of column & number of row*. For instance, the MST string "Yes" (see Table II) is linked with its fully-specified name "Traversed Esophageal Stenosis" with the location attribute: *6&3&8*.

Table II: Two examples of unspecified MST terms and fully specified MTHMST terms.

- "Malignant intrinsic" → *Malignant intrinsic Appearance Stenosis Esophageal Lumen* → **Malignant intrinsic Appearance of Esophageal Stenosis**
- "Yes" → *Yes Traversed Stenosis Esophageal Lumen* → Traversed Esophageal Stenosis

Semantic Types Assignment (categorization)

A default STYs assignment, according to the intended meaning of the MST table titles, proved not to be useful since there is a huge amount of semantic heterogeneity within the tables. The STYs were then assigned manually by the primary author, trying to retain consistency and to follow the intended meaning of the STY definitions. After insertion in the Metathesaurus database (see below) further editing of the proposed STYs was necessary, since the algorithms used by Apelon resulted in the mapping of MTHMST2001 atoms to existing Metathesaurus concepts, and this made explicit a number of questionable automatic assignments.

A second STY, "Therapeutic Procedure", was assigned to some high level terms (e.g. Colonoscopy) originally categorized with just "Diagnostic Procedure", but in general assignment of multiple STYs was avoided, to minimize the risk of concept-level ambiguity.

MST_Classes Assignment

In order to retain as much as possible of the intended meaning of the source as expressed with the general organization of MST tables, a source-level attribute ("MST_Class") was added. MST_Classes were named according to the names of the table title and column titles, separated by periods. A list of Classes,

fully specified names and Semantic Types for a sample of Table no. 14 (Additional Procedures) terms is reported in Table III.

Table III: MST_Classes, MTHMST2001 Name and UMLS Semantic Types (STY) for a sample of four terms (Table 14 of MST).

MST_Class	Name	STY
Procedure.Heading	Diagnostic Procedure	Diagnostic Procedure
Procedure.Term	Biopsy	Diagnostic Procedure
Procedure.Attribute	Biopsy device	Medical Device
Procedure.Attribute.Value	Biopsy forceps	Medical Device

Concept and source-level relationships and their attributes

The UMLS SN was extensively checked for appropriate relationship attributes (RelA), or names of relations, to be assigned to concept-level relationships between MST terms and existing Meta concepts or new Meta concepts introduced with the MST integration.

Tentative RelA assignment for a number of MTHMST2001 atoms and concepts was not possible without altering the intended organization of the source, or without misusing the SN rules that specify allowable STY-RelA-STY structures[7]. Source-level relationships not derived from the SN were added for these concepts; they apply only to MTHMST2001 atoms and are not allowed at a conceptual level.

Inter-table relationships assignment

Explicit inter-table relationships (all with RelA "has_location") were algorithmically derived for the "site(s)" notations related to Findings, Details of the Procedure and Additional Procedures sections. These relations were then edited by the primary author in order to prevent the occurrence of semantic inconsistencies and unallowable relations.

3) Insertion in the Metathesaurus database

The mapping of the MST schema into the Meta schema of the DB prepared during the inversion process was then used to insert the atoms and relations into the Metathesaurus. A manual control of the insertion product did not show major technical problems.

4) Editing and QA review of the process

Worklists of concepts were prepared and reviewed extensively, resolving missed synonymy, refining STY assignment, and checking the consistency of relationships. Documentation of the inversion process was prepared for reference.

RESULTS

During the inversion process a detailed analysis of the MST has been obtained, and this allowed for the production of a new source (MTHMST2001 for "Metathesaurus-MST, release 2001"). A comparison of the structure and content of MTHMST2001, and of the MST itself, with authoritative recommended criteria for vocabulary production and integration[8,9] is now feasible. A strengthening of the informational model of MST was obtained: while the original information has been thoroughly retained, a more explicit internal structure has emerged. During MST integration in UMLS:

1) a list of unique fully-specified terms was produced. This is a prerequisite for enabling the MST content to be machine-readable

2) the terms were assigned to the concept-level structure of the Metathesaurus

3) to all new concepts added to UMLS appropriate Semantic Types (STY) were assigned, and STY of existing concepts were reviewed and, where necessary, edited in an effort to fix previous inadequate categorization. MTHMST2001 concepts now inherit properties, hierarchies and relationships from the vast conceptual space of UMLS

4) complete sets of concept-level and source-level explicit relationships among the terms have been produced. Each term can now be mapped to both the UMLS SN and to an explicit representation network which, while embedded in the original source, was not machine-readable in it.

DISCUSSION

Problems encountered

A number of problems in the original MST tables were found: typographical errors, different granularity of the three national language versions in some areas, subtle inconsistencies of organization of terms within the tables. Appropriate solutions were applied.

For Italian and French versions of the terms, the problem of representing diacritical characters will need to wait for a UNICODE-compatible release of UMLS for a satisfactory solution.

The meronymies (*part_of* hierarchies) derived from the first group of tables ("sites") seem to be heterogeneous. Some hierarchical relationships also are heterogeneous across, and also within, the tables. A more robust and principled analysis of the MST could be useful to enhance the terminology with a real "ontological layer" expressive enough to support reasoning and resource inter-operability, but for this

project it was judged more important to respect the source organization, retaining the general informational model of MST.

While in general relationships borrowed from the UMLS SN were enough to represent the source-level relations, peculiar representation problems needed a number of specific relationships, and this is the reason for some relationships left undefined at source level.

A practical issue is related to the *roles of terms* in the informational model of GIE domain. Of the 1945 atoms in MTHMST2001, 290 (15%) are clustered in multiple-atom concepts. While 237 of them are allowed synonyms, the remaining 53 (2.7% of total number of atoms) are merged in the same concept but used in different roles within MST: 47 instances of Finding/Diagnosis, 3 of Reason/Diagnosis (for example: "Stent occlusion as reason for ERCP" and "Stent occlusion as main diagnosis for the biliary tract" were merged in a single concept, with STY: Disease or Syndrome), 2 of Finding/Site and even 1 instance of Reason/Procedure.

In Cimino's words[10] these are, at least, context-dependent ambiguous terms, as opposed to context-independent (concept level or true) ambiguous terms. The Metathesaurus could possibly be enhanced with a dedicated attribute for the specification of the *role*, or *intended use*, of source's, or source's derived, terms.

Future plans

An application-oriented but application-independent ontology for the domain of gastrointestinal endoscopy can be developed on top of the UMLS-enabled MST.

Formal mapping of MST "site" terms with the anatomical terms of UWDA[10] is under way to test to what extent properties of the UWDA are inheritable in the framework of this application-oriented ontology.

The MST Committee efforts are now focused on adding textual and symbolic (canonical images) definitions to the terms. These could be easily added to the UMLS-enabled MST, also as a reference for the formal definition of the ontology layer.

CONCLUSION

Integration of a standard medical subdomain (GI endoscopy) terminology in the UMLS led to a fairly profound enhancement of the terminology. The UMLS Metathesaurus, 2002 edition, now hosts both all the terms of the original MST and their fully specified and categorized MTHMST2001 equivalents, with a full set of explicit relationships. While mapping of MST to other representational schema have been proposed[11], we argue that our integration of the internationally validated non-proprietary MST in the UMLS is better suited to enhance the interoperability of clinical databases relevant to outcomes research in gastroenterology.

Acknowledgements

Giuliana, Valeria, Agnese and Teresa warmly supported MT in his sabbatical at NLM. We are in debt to Olivier Bodenreider, Anita Burgun, Thomas Rindflesch, Guy Divita for counselling and support. William King and Robert Hawks of Apelon, Inc. assisted during inversion and Laura Roth and Tammy Powell at NLM helped with the insertion of the data into the Metathesaurus.

REFERENCES

[1] Campbell KE, Musen MA: Creation of a systematic domain for medical care: the need for a comprehensive patient-description vocabulary. MEDINFO 92. 1992 Elsevier Science Publishers B.V. 1437-42.

[2] UMLS Knowledge Sources. 12th ed., January 2001. National Library of Medicine, Bethesda, USA.

[3] MST© Copyright 1995 European Society of Gastrointestinal Endoscopy (ESGE). Available at: http://www.omed.org/minimal.htm Accessed on March 21, 2001.

[4] Delvaux M, Crespi M, Armengol-Miro JR, et al: Minimal standard terminology for digestive endoscopy: results of prospective testing and validation in the GASTER project. Endoscopy. 2000 Apr;32(4):345-55

[5] Logan JR, Klopfer KC: The use of a standardized terminology for comparison of free text and structured data entry. Proc AMIA Symp. 2000:512-6.

[6] McCray AT, Bodenreider O, Malley JD, Browne AC: Evaluating UMLS Strings for Natural Language Processing. Proc. AMIA Symp. 2001:448-52.

[7] UMLS Knowledge Source Server. Available at: http://umlsks.nlm.nih.gov/ Accessed on January 2002

[8] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med. 1998 Nov;37(4-5):394-403. Review.

[9] Rector A: Clinical terminology: Why Is it so Hard? Method Inform Med 1999; 38:239-52.

[10] Mejino JL Jr, Rosse C. The potential of the digital anatomist foundational model for assuring consistency in UMLS sources. Proc AMIA Symp. 1998:825-9.

[11] Korman LY, Bidgood WD Jr. Representation of the Gastrointestinal Endoscopy Minimal Standard Terminology in the SNOMED DICOM microglossary. Proc AMIA Annu Fall Symp. 1997:434-8.