Journal of Biomedical Informatics 43 (2010) 988–997

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Source authenticity in the UMLS – A case study of the Minimal Standard Terminology

Gai Elhanan [*], Kuo-Chuan Huang, Yehoshua Perl

New Jersey Institute of Technology, Newark, NJ, USA

ARTICLE INFO

Article history:
Received 22 April 2010
Available online 6 August 2010

Keywords:
UMLS
MST
Terminology
Endoscopy
Gastrointestinal
Integration

ABSTRACT

As the UMLS integrates multiple source vocabularies, the integration process requires that certain adaptation be applied to the source. Our interest is in examining the relationship between the UMLS representation of a source vocabulary and the source vocabulary itself. We investigated the integration of the Minimal Standard Terminology (MST) into the UMLS in order to examine how close its UMLS representation is to the source MST. The MST was conceived as a "minimal" list of terms and structure intended for use within computer systems to facilitate standardized reporting of gastrointestinal endoscopic examinations. Although the MST has an overall schema and implied relationship structure, many of the UMLS integrated MST terms were found to be hierarchically orphaned, and with lateral relationships that do not closely adhere to the source MST. Thus, the MST representation within the UMLS significantly differs from that of the source MST. These representation discrepancies may affect the usability of the MST representation in the UMLS for knowledge acquisition. Furthermore, they pose a problem from the perspective of application developers. While these findings may not necessarily apply to other source terminologies, they highlight the conflict between preservation of authentic concept orientation and the UMLS overall desire to provide fully specified names for all source terms.

1. Introduction

The stated purpose of the UMLS [1–4] is to "facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health" [1]. The Metathesaurus (META) [2,5–7] is "a very large . . . vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them" [2]. To achieve its purpose the UMLS integrates multiple source vocabularies into its structure. This is not an easy task as many source vocabularies overlap in their coverage, but with each source modeled differently, and oriented towards different tasks. As a result, the integration process requires certain adaptations during the incorporation of source terminologies. Creation of fully specified names, mapping of terms and synonyms to existing concepts, granularity resolution, interpretation and addition of hierarchical and lateral relationships, and assignment of semantic types of the Semantic Network [2,8–10] are among the significant tasks in the process [11–15].

Thus, despite the flexible structure of the META, these adaptations may result in a source image within the META that is not truly authentic compared to the source vocabulary. For illustration purposes we have chosen the Minimal Standard Terminology (MST) which was integrated in 2002 into the META [11], and is one of the few UMLS source terminologies, the integration of which was reported. The MST was conceived as a "minimal" list of terms and structure to be used within computer systems to facilitate standardized reporting of gastrointestinal endoscopic examinations (GIE) [16]. It consists of 24 interlinked tables, each containing terms, qualifiers and modifiers that cover 95% of the commonly used terms in GIE reports. The MST is a well encapsulated, domain-specific terminology and therefore makes a good candidate for assessment.

We present a qualitative examination of the integration of the MST in the UMLS META with respect to the structure of the MST, the resulting knowledge representation, and the ability of the MST integration to support applications according to the original design of the MST and the original intent of the integration as stated in [11].

2. Background

The development of electronic endoscopes and the emergence of large-scale hospital information systems (HIS) in the early 1990s brought to light the need to disseminate the information collected during such examinations in a standardized, searchable

[*] Corresponding author. Address: Computer Science Department, New Jersey Institute of Technology, University Heights, Newark, NJ 07102-1982, USA.
E-mail address: gelhanan@gmail.com (G. Elhanan).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.07.008

and sharable manner [16]. The European Society for Gastrointestinal Endoscopy (ESGE), along with the American Society for Gastrointestinal Endoscopy (ASGE) and several other entities embarked on the creation of such standardized terminology. The resulting Minimal Standard Terminology (MST) was conceived as a "minimal" list of terms to be used within computer systems to facilitate standardized reporting of gastrointestinal endoscopic examinations (GIE). As opposed to the terminology developed by the World Organization of Digestive Endoscopy (OMED) [17,18], the MST is not a full fledged terminology but more of a controlled collection of preferred terms structured in a manner that follows the logical input and collection of information related to gastrointestinal endoscopy [16,19,20]. The MST was not designed to be an all-comprehensive collection but was meant to cover the most commonly used and widely acceptable items.

MST (version 2) was incorporated into the UMLS January 2002 release [11]. The drive for the importation of the MST into the UMLS Metathesaurus (META) was to map MST terms into the META to provide a new machine-readable, MST-compatible terminological tool.

The UMLS META "reflects and preserves the meanings, concept names, and relationships" from its source vocabularies [2]. Terms from various sources are mapped into concepts, and relationships between terms are mapped into relationships between the corresponding concepts. META relationships are defined as triples consisting of a relationship type (REL), a source concept, and a target concept. These mappings are based primarily on the sources, but the UMLS integration team may make adjustments as deemed appropriate. Intra-source relationships are explicit or implied in the original source vocabulary and represent its hierarchical order and cross-reference structure. Inter-source relationships are introduced by UMLS editors during content integration of a source to connect otherwise orphaned concepts or isolated components from the new source to the rich information content of other sources already existing in the META. Those relationships that are presented as hierarchical in the source are recorded as parent–child (PAR–CHD) [21]. Relationships deemed as hierarchical by the UMLS editors are recorded as broader–narrower (RB–RN) [22]. Also included are lateral (non-hierarchical) relationships. META relationships are further qualified by a label, called a relationship attribute (RELA). These RELAs are obtained from source vocabularies directly and further explain the nature of the relationship beyond the wide relationship-type categories. RELAs can be assigned to both hierarchical and lateral RELs.

The MST is organized in a set of 24 tables following four main segments: examination sites, reasons for the examination, examination specific data and organ specific information. The organ specific information is further divided into three sub-segments: findings, additional procedures and diagnoses. The MST has an overarching schema as depicted in Fig. 1 (see also [16]). However, this schema is not explicitly provided with MST's data, but rather implied through the relationship between tables and the table structure itself.

An example of a MST table is depicted in Table 1 which lists the terms to be used when describing observations made during examination of the pancreas. The two-dimensional structure of each MST table represents the lower level information model of the GIE record. Each table has up to five columns: Headings, Terms, Attributes, Attribute Values and Sites. The Headings column further segments each table into general classes specific to the MST

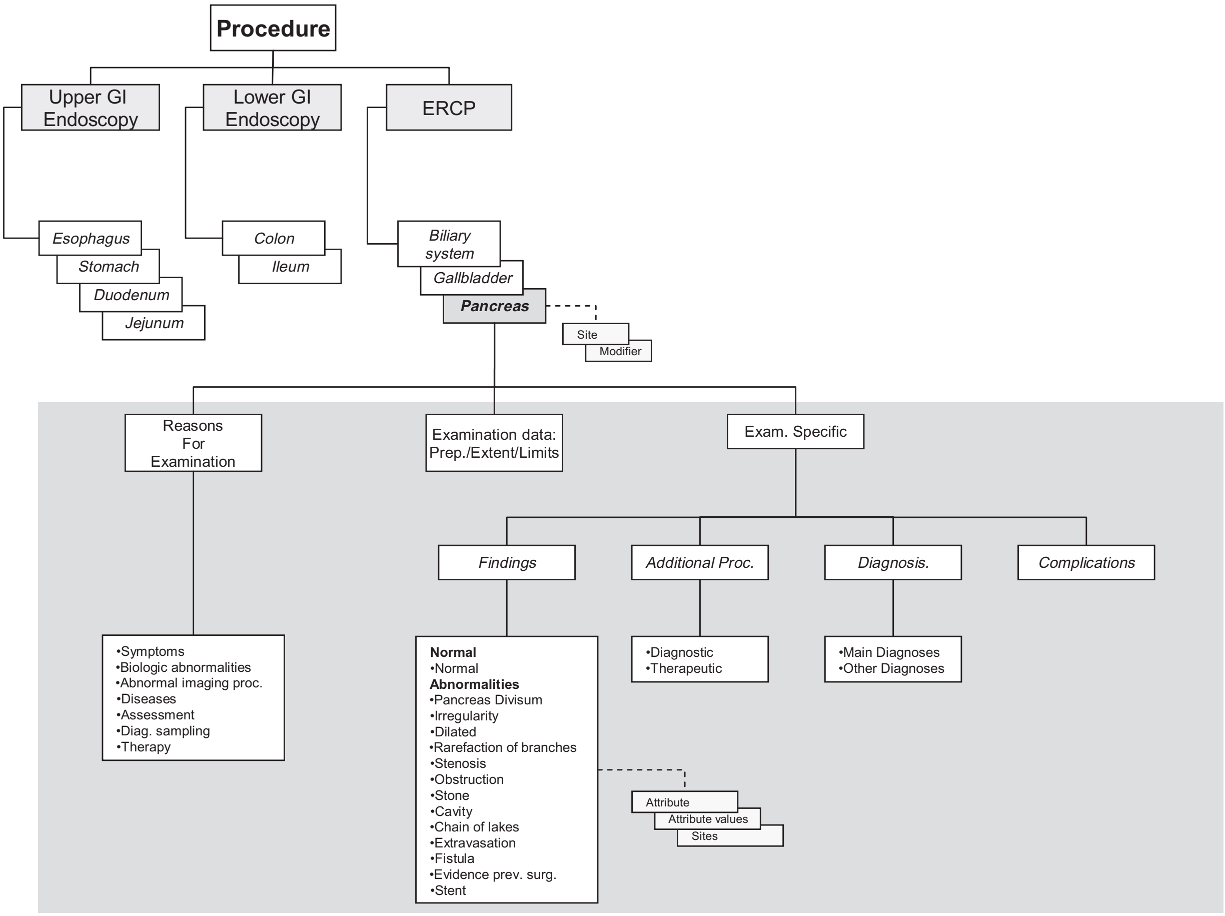


Fig. 1. MST organization: tree structure of terms for endoscopic findings. The ERCP pancreatic component is exploded (shaded, some terms and attributes are not displayed).

Table 1
List of terms used to describe observations during ERCP examination of the pancreas (MST Table 12).

Headings	Terms	Attributes	Attributes values	Sites
Normal	Normal			
Abnormalities				
	Pancreas Divisum	Extent	Complete Incomplete	
	Irregularity	Extent	Localized Segmental Diffuse	Site(s)
	Dilated	Extent	Localized Segmental Diffuse	Site(s)
	Rarefaction (attenuation) of branches	Extent	Localized Segmental Diffuse	Site(s)
	Stenosis	Length	In mm	Site(s)
	Obstruction	Appearance	Stone Tumor/ mass	Site(s)
		Completeness	Partial Complete	
	Stone	Number	Single Multiple	Site(s)
	Cavity	Number	Single Multiple	Site(s)
		Diameter	In mm	
	Chain of lakes			Site(s)
	Extravasation			Site(s)
	Fistula			Site(s)
	Evidence of previous surgery			Site(s)
	Stent			Site(s)

segment or sub-segment the table represents. As depicted in Table 1, under the **Abnormalities** heading (column 1), the term **Pancreas Divisum** (column 2) requires an attribute Extent (column 3) with two possible attribute values, **Complete** and **Incomplete** (column 4). The table structure also indicates that a site specification is not required for the term **Pancreas Divisum**. Overall, the MST covers 8 endoscopic procedure, 93 anatomical sites, 122 reasons, 52 GIE specific data elements, 1030 findings, 166 additional procedures, 235 diagnoses and 7 complications, totaling 1713 unique data elements.

As noted by Tringali et al. [11], the MST lacks fully specified terms but rather requires the user to post-coordinate column values based on the constraints presented by each MST table structure and content. Such organization lends itself well when used for structured or semi-structured data entry or retrieval by humans or computer programs. However, the UMLS META representation of the MST is based on fully specified terms and therefore for successful yet MST-compatible integration, all pertinent post-coordination needed to occur beforehand. Hence fully specified names had to be created with appropriate assignment of Semantic Types (ST) from the UMLS Semantic Network (SN). The fully specified terms thus created in META must be linked back to their respective MST tables, attributes and attribute values. At the end of the process, the 2002 version of the UMLS META contained 1945 MST-related terms (source-abbreviated as MTHMST) with a full set of explicit relationships that can be used for interoperability between GIE clinical repositories.

3. Methods

Below is a qualitative examination of the integration of the MST in the UMLS META. The analysis was performed using the UMLS 2007 AC release. As a reference, the published MST version 2.0 h Fixed (March 9, 1999) was used. All MST-related concepts were identified based on the MST source (MTHMST). Semantic types, relationships types and relationship attributes of the relevant META concepts were identified, as were parents and children within the UMLS hierarchy.

3.1. Concept-focused analysis

META-mapped concepts were evaluated for the following:

- Placements within relevant UMLS hierarchy based on parent (PAR) or child (CHD) relationships.
- Availability of broader (RB) or narrower (RN) relationship types as indicators of hierarchical relationships.
- Other (RO) relationship types.
- ST assignments.

3.2. Structural analysis

Specific attention was given to the pancreas-related Endoscopic Retrograde Cholangiopancreatography (ERCP) tables within the MST (Tables 3, 5, 12, 18 and 23). The conceptual representation of the MTHMST META terms was used in an attempt to re-recreate the original MST ERCP tables' structure and content. The resulting recreated structures were compared to the respective original MST tables and discrepancies were recorded and analyzed.

4. Results

Each original MST term was mapped based on meaning and UMLS MTHMST source to META concepts. Overall, 1945 MST terms mapped to 1636 UMLS concepts of which, 1305 were newly introduced from the MST (80% of MST concepts).

4.1. Hierarchical placement

Many MST-derived terms and their associated concepts were found to be isolated within the UMLS. We note such concepts as singletons: i.e. concepts that are not part of a larger group of concepts based on the pattern of their outgoing hierarchical or lateral relationships. We observed three major categories of such singletons. We note that these singleton categories are non-disjoint:

(1) UMLS singletons: MST-derived concepts without any hierarchical relationships (PAR, CHD, RB, RN) in the UMLS.
210 such concepts were identified.
Example: The MST-derived concept **Duct cannulation result** (C0939994) is a UMLS singleton since it has no assigned outgoing hierarchical relationships (PAR, CHD, RB, RN) in the MST or any other UMLS source vocabulary.

(2) MST singletons: MST-derived terms and concepts without any hierarchical relationships (PAR, CHD, RB, RN) in the MTHMST component of the UMLS.
286 terms and 272 such concepts were identified.
Example: The MST term **Balloon** (A0652270), mapped to the UMLS concept **Balloon Dilatation** (C0004704), has many hierarchical relationships from multiple UMLS sources but only an RO (*uses*) relationship derived from the MTHMST to **Dilatation** (A0012207) (Table 2).
Many of the concepts and terms in this category overlap with concepts in the previous category.

(3) MST super-singletons: MST-derived terms and concepts without any relationships (PAR, CHD, RB, RN, RO) in the MTHMST component of the UMLS.
65 terms and 61 such concepts were identified.

Example: The MST term **Acute pancreatitis** (A0242598) was mapped to the UMLS concept **Acute pancreatitis unspecified** (C0001339). Across all UMLS source terminologies the concept has numerous PAR, CHD, RB, RN and RO relationships. However, within the MTHMST UMLS subset, the term lacks any relationship altogether and is completely isolated (Table 2).

Many of the terms in this category overlap with the concepts and terms in the previous category.

All the hierarchical relationships found to exist in the MTHMST are of the RB/RN type with no PAR or CHD relationships incorporated from the MST. Of the RB/RN hierarchical relationships, 1279 use the *isa*/*inverse_isa* attribute and 97 use other RB/RN attributes such as *has_part*/*part_of*. No MST-related concepts were found to completely lack any relationships within the META.

4.2. Recreation of the MST ERCP-related tables

Utilizing the META hierarchy as well as MTHMST RN/RB relationships and other relationships to re-recreate MST ERCP-related table structure and content (Tables 3, 5, 12, 18 and 23 of MST version 2.0) demonstrated the following:

- The overarching schema of the MST (see Fig. 1) is not conceptualized in the META. While some of the relations between MST tables are depicted as lateral relationships, it is an incomplete bottom-up representation. The UMLS does not contain any top-down, comprehensive representation of the actual structure of the MST and the complex association between its tables.
- The actual use of META relationship types and attributes to model the relationships between MST table and table column content is not self evident or consistent. The utilized UMLS relationship types and attributes do not closely resemble the implied and explicit relationships of the MST.
- The use of relationship attributes in MTHMST is incomplete and the attribute descriptions do not necessarily resemble the MST relevant relationship.
- Manual re-modeling of the MST based on the MTHMST content of META is difficult and requires extensive intrinsic knowledge of the published MST.

Figs. 2–4 demonstrate different levels of the above mentioned points. In these figures an attempt was made to follow the MST schema of GIE (Fig. 1) to re-reproduce sections of different MST tables based on the MTHMST META source. At the top of each figure is the relevant table data of the MST and below is the related, reconstructed data based on the UMLS.

Table 2
PAR, CHD, RB, RN and RO relationships for C0004704 and C0001339 from all UMLS sources and from MTHMST.

All sources relationships	MTHMST relationships
Balloon Dilatation (C0004704) [Balloon/A0652270/MST]	
• Parents (18)	• Other (1)
• Children (12)	
• Broader (4)	
• Narrower (17)	
• Other (43)	
Acute pancreatitis unspecified (C0001339) [Acute pancreatitis/A0242598/MST]	
• Parents (32)	None
• Children (58)	
• Broader (9)	
• Narrower (43)	
• Other (70)	

4.2.1. Lack of MST schema conceptualization

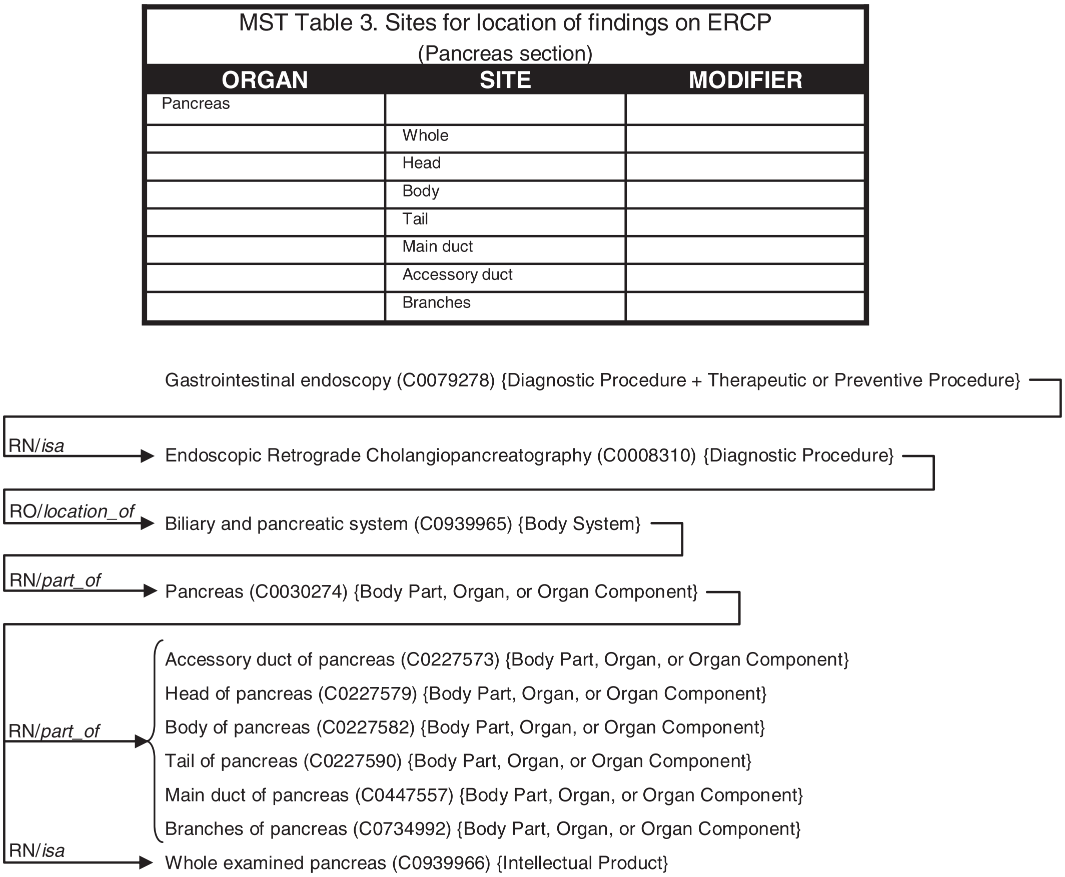
In Fig. 2, a path via various relationships can be created to represent the Pancreas section of MST Table 3. However, to achieve this, explicit knowledge of the relationship types and specific attributes at each level of the path is required. This information is not modeled in the META and can only be obtained by simultaneous use of the published MST tables. For example, in order to create the set of MST terms for pancreatic sites of ERCP findings, one has to know to combine the RN/*part_of* and the RN/*isa* for a complete result set. Fig. 3 demonstrates thee flattening effect of the UMLS representation of parts of the Attribute and Attribute value columns of MST Table 5 as shown for the Result, Method and Device attributes. While these attributes exist in the UMLS, they lack any relationship to their respective attribute values. Fig. 4 demonstrates that there is no top-to-bottom path between the individual reasons for performing ERCP and the ERCP examination or any GIE examination. While the **ERCP** (C0008310) and **Reasons for gastrointestinal endoscopy** (C090011) concepts exist as MTHMST terms, they have no relationships to the individual reasons, many of which exist as super-singletons. The modeling of the diseases attributes in MST Table 18 is also not according to the source MST. While it is clear in the printed version of MST that any of the Attributes should apply to any of the Diseases in Table 18 (Fig. 4), the MTHMST representation is independent and does not provide any information regarding that relationship. Thus, the list of Diseases is completely separated from the need to qualify each one by the Attributes (Suspected, Established, etc.).

4.2.2. Semantics mismatch

The mapping of MST terms into existing or newly created META concepts within the context of the MST tables raises many issues regarding Semantic Types (STs). For example, in Fig. 2 the MST concept **Whole examined pancreas** (C0939966) has the ST of {Intellectual Product}. This concept has an *isa* relationship with **Pancreas** which has the {Body Part, Organ, or Organ Component} ST. The two are clearly not related within the UMLS Semantic Network (SN) tree. In Table 1 the concept **Pancreas divisum** (C0266270) is part of MST Table 12 which lists all possible findings of an examination of the pancreas. This implies a {Finding} ST but within the META **Pancreas divisum** has the {Congenital Abnormality} ST. The two semantic types reside in two distinct branches of the SN.

In some cases the mapping of MST terms to existing META concepts coupled with the generation of fully specified names seem to have resulted in semantic inconsistencies. For example, the MST attribute value **Balloon** (A0652270) in MST Table 14 (Fig. 5), was mapped to **Balloon Dilatation** (C0004704) with a {Therapeutic or Preventive Procedure} ST. However, the original MST term was **Balloon** and, along with **Guided bougie** (C0941222) and **Non-guided bougie** (C0940881), provide the qualifiers for the type of dilatation. **Guided bougie** and **Non-guided bougie** were mapped to existing identical concepts with a ST of {Medical Device}, while **Balloon** received a procedural ST which is inconsistent with the other two members of the set as well as with the RO/*used_by* relationship type and attribute assigned to all three. Such inconsistencies also exist in the treatment of identical MST attributes. Examine the example in Fig. 5: as described above, the various dilatation types are linked to **Dilatation** utilizing a RO/*used_by* relationship type and attribute. However, in the same table the various types of **Thermal Therapy** (A1999492) procedures utilize an RN/*isa* relationship type and attribute for the same MST attribute Type.

As with the **Dilatation** example above, the only retained link to the original MST table column headings is preserved via the assigned UMLS META Attribute Name (ATN). In this case, the **Dilation** (A0012207) atom has the assigned value of Minimal Standard (Terminology) Class (MSC) {Procedure.Term} [AT18102260/MTHMST] while **Type of dilation** (A1999538) was assigned MSC



MST Table 3. Sites for location of findings on ERCP
(Pancreas section)

ORGAN	SITE	MODIFIER
Pancreas		
	Whole	
	Head	
	Body	
	Tail	
	Main duct	
	Accessory duct	
	Branches	

Gastrointestinal endoscopy (C0079278) {Diagnostic Procedure + Therapeutic or Preventive Procedure}
RN/*isa*
Endoscopic Retrograde Cholangiopancreatography (C0008310) {Diagnostic Procedure}
RO/*location_of*
Biliary and pancreatic system (C0939965) {Body System}
RN/*part_of*
Pancreas (C0030274) {Body Part, Organ, or Organ Component}
RN/*part_of*
Accessory duct of pancreas (C0227573) {Body Part, Organ, or Organ Component}
Head of pancreas (C0227579) {Body Part, Organ, or Organ Component}
Body of pancreas (C0227582) {Body Part, Organ, or Organ Component}
Tail of pancreas (C0227590) {Body Part, Organ, or Organ Component}
Main duct of pancreas (C0447557) {Body Part, Organ, or Organ Component}
Branches of pancreas (C0734992) {Body Part, Organ, or Organ Component}
RN/*isa*
Whole examined pancreas (C0939966) {Intellectual Product}

Fig. 2. On top is the Pancreas section from MST Table 3 – site for location of findings on ERCP. The bottom section represents an effort to recreate the content of Table 3 utilizing available MTHMST hierarchical and other relationships. Relationship type and attribute are marked above the arrows. Concepts are presented by preferred name (Concept ID) {semantic type}.

{Procedure.Attribute} [AT18102369/MTHMST] and the **Guided bougie** (A1998619) atom has MSC {Procedure.Attribute.Value} [AT18102868/MTHMST]. As demonstrated, the META Attribute Names correctly represent the column heading in this example from MST Table 14 (Fig. 5). However, the inter-relationship or hierarchy between the META Attribute Names is not captured anywhere else and therefore requires external information regarding the MST to help resolve cases of flattened relationships.

Tringali et al. [11] also describe another mechanism to map fully specified terms to their original unspecified strings. This mechanism retains the relative location of the unspecified strings within the table/column/row structure of the MST in the format of *number of table and number of column and number of row*. However, an exhaustive search of UMLS 2007 AC release did not uncover this feature as part of the data set and it is also not present in subsequent UMLS releases.

4.2.3. Authenticity of structure

The example of thermal therapy types also demonstrates the difficulty recreating the exact data set per MST table. In MST Table 14 **Thermal Therapy** as an additional diagnostic and therapeutic procedure, is allowed only two types. Utilizing the available MTHMST RN/*isa* relationship generates a set of five procedures (Fig. 5); two for the Type attribute and three for the Purpose attribute for **Thermal Therapy**, without any information that will enable to differentiate between them. Thus, the MTHMST does not adhere to the original MST structure. While keeping most of the attributes, attribute values are linked directly to MST table terms via inconsistent relationship types and relationship attributes. It is also noteworthy that the purpose attribute is missing altogether.

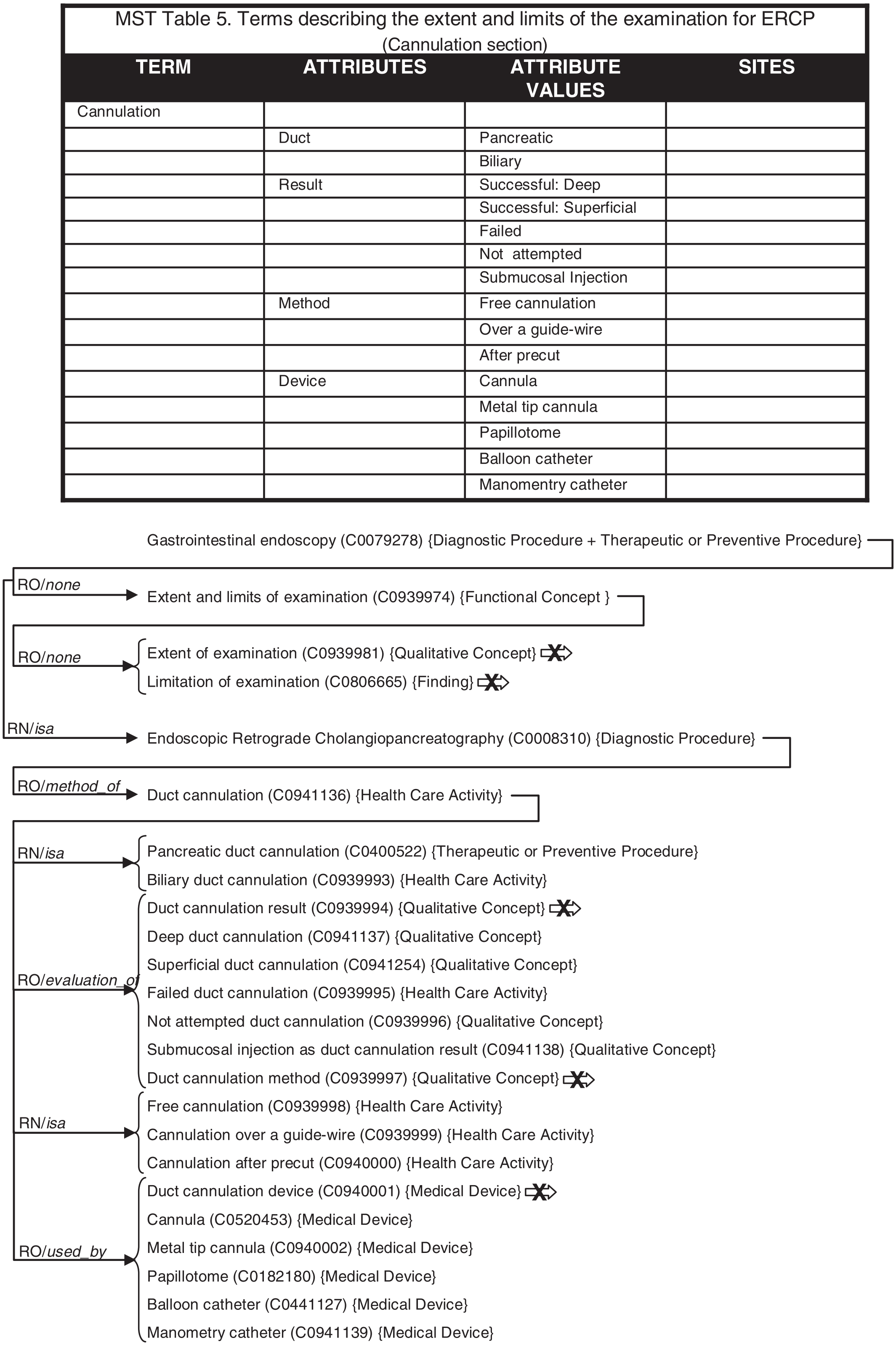
Conversely, there are other cases where the data set provided in the UMLS via relationships is incomplete. MST Table 12 (Table 1) lists 13 pancreatic abnormalities under the Terms column. However, in MTHMST **Pancreatic abnormality** (A1999024/C0940732) has only 12 narrower concepts (RN/*isa*) and no other related concepts. The MST term **Stenosis** as MTHMST's **Pancreatic duct stenosis** (A2015640/C0940747) is not linked by any relationship to **Pancreatic abnormality** or to any other MTHMST term, and stands as a super-singleton in the MTHMST.

5. Discussion

5.1. Significance

The UMLS is an important system that enables mapping between source vocabularies thus promoting potential interoperability between systems that utilize different sources of the UMLS. In order to offer functionality beyond simple concept mapping, the UMLS must also retain additional information related to concept definitions. That information is usually represented by the network of relationships assigned to the source terminology concept, their meaning and inter-relationships with other concepts. The UMLS framework, however, may not be fully (conceptually and functionally) compliant with source vocabularies design and various degrees of adaptation may be required to successfully provide a functional integration.

The sheer size of many of the UMLS source terminologies renders *post hoc* examination of their integration impractical. The MST however, was designed to be a practical, encapsulated, light weight GIE terminology, in part in response to difficulties with



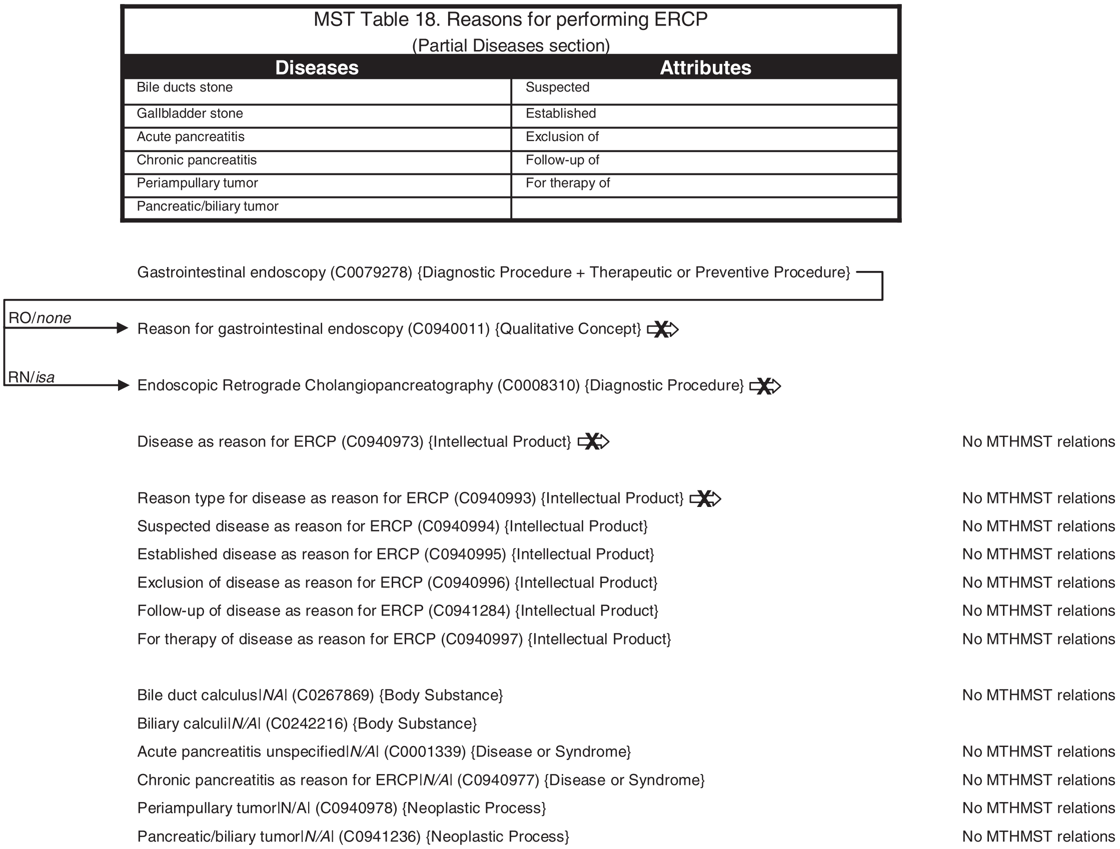
MST Table 5. Terms describing the extent and limits of the examination for ERCP (Cannulation section)

TERM	ATTRIBUTES	ATTRIBUTE VALUES	SITES
Cannulation			
	Duct	Pancreatic	
		Biliary	
	Result	Successful: Deep	
		Successful: Superficial	
		Failed	
		Not attempted	
		Submucosal Injection	
	Method	Free cannulation	
		Over a guide-wire	
		After precut	
	Device	Cannula	
		Metal tip cannula	
		Papillotome	
		Balloon catheter	
		Manomentry catheter	

Gastrointestinal endoscopy (C0079278) {Diagnostic Procedure + Therapeutic or Preventive Procedure}
RO/none
Extent and limits of examination (C0939974) {Functional Concept }
RO/none
Extent of examination (C0939981) {Qualitative Concept} ⇏
Limitation of examination (C0806665) {Finding} ⇏
RN/isa
Endoscopic Retrograde Cholangiopancreatography (C0008310) {Diagnostic Procedure}
RO/method_of
Duct cannulation (C0941136) {Health Care Activity}
RN/isa
Pancreatic duct cannulation (C0400522) {Therapeutic or Preventive Procedure}
Biliary duct cannulation (C0939993) {Health Care Activity}
Duct cannulation result (C0939994) {Qualitative Concept} ⇏
Deep duct cannulation (C0941137) {Qualitative Concept}
Superficial duct cannulation (C0941254) {Qualitative Concept}
RO/evaluation_of
Failed duct cannulation (C0939995) {Health Care Activity}
Not attempted duct cannulation (C0939996) {Qualitative Concept}
Submucosal injection as duct cannulation result (C0941138) {Qualitative Concept}
Duct cannulation method (C0939997) {Qualitative Concept} ⇏
Free cannulation (C0939998) {Health Care Activity}
RN/isa
Cannulation over a guide-wire (C0939999) {Health Care Activity}
Cannulation after precut (C0940000) {Health Care Activity}
Duct cannulation device (C0940001) {Medical Device} ⇏
Cannula (C0520453) {Medical Device}
Metal tip cannula (C0940002) {Medical Device}
RO/used_by
Papillotome (C0182180) {Medical Device}
Balloon catheter (C0441127) {Medical Device}
Manometry catheter (C0941139) {Medical Device}

Fig. 3. On top is the Cannulation section from MST Table 5 – terms describing the extent and limits of the examination for ERCP. The bottom section represents an effort to recreate the content of Table 5 utilizing available MTHMST hierarchical and other relationships. Relationship type and attribute are marked above the arrows. Concepts are presented by preferred name (Concept ID) {semantic type}. ⇏ indicates table attributes concepts that lack relationships to their respective table attribute values.

the use of more complicated and comprehensive data sets [16–18,20]. In order to facilitate data sharing, the MST also contains information about the structure and the required elements to record GIE exams. The MST UMLS integration effort was designed to provide a MST-compatible, machine-readable terminological tool [11]. Thus, the MST UMLS integration offers a well encapsu-



MST Table 18. Reasons for performing ERCP
(Partial Diseases section)

Diseases	Attributes
Bile ducts stone	Suspected
Gallbladder stone	Established
Acute pancreatitis	Exclusion of
Chronic pancreatitis	Follow-up of
Periampullary tumor	For therapy of
Pancreatic/biliary tumor	

Gastrointestinal endoscopy (C0079278) {Diagnostic Procedure + Therapeutic or Preventive Procedure}
RO/none
Reason for gastrointestinal endoscopy (C0940011) {Qualitative Concept}
RN/isa
Endoscopic Retrograde Cholangiopancreatography (C0008310) {Diagnostic Procedure}
Disease as reason for ERCP (C0940973) {Intellectual Product}
No MTHMST relations
Reason type for disease as reason for ERCP (C0940993) {Intellectual Product}
No MTHMST relations
Suspected disease as reason for ERCP (C0940994) {Intellectual Product}
No MTHMST relations
Established disease as reason for ERCP (C0940995) {Intellectual Product}
No MTHMST relations
Exclusion of disease as reason for ERCP (C0940996) {Intellectual Product}
No MTHMST relations
Follow-up of disease as reason for ERCP (C0941284) {Intellectual Product}
No MTHMST relations
For therapy of disease as reason for ERCP (C0940997) {Intellectual Product}
No MTHMST relations
Bile duct calculus|NA| (C0267869) {Body Substance}
No MTHMST relations
Biliary calculi|N/A| (C0242216) {Body Substance}
Acute pancreatitis unspecified|N/A| (C0001339) {Disease or Syndrome}
No MTHMST relations
Chronic pancreatitis as reason for ERCP|N/A| (C0940977) {Disease or Syndrome}
No MTHMST relations
Periampullary tumor|N/A| (C0940978) {Neoplastic Process}
No MTHMST relations
Pancreatic/biliary tumor|N/A| (C0941236) {Neoplastic Process}
No MTHMST relations

Fig. 4. On top is part of the Diseases section from MST Table 18 – reasons for performing ERCP. The bottom section represents an effort to recreate the content of Table 18 utilizing available MTHMST hierarchical and other relationships. Relationship type and attribute are marked above the arrows. Concepts are presented by preferred name (Concept ID) {semantic type}. ✗ indicates concepts that lack relationships to their respective diseases and attributes column content.

lated and manageable sub-domain to examine the integration effort and its compatibility with the original source terminology.

The MST published structure is not organized in a fashion that might be expected from other UMLS source terminologies. Hierarchical and other relationships are implied through table and column names and must be derived by the user. The MST table design defines the allowed interactions between terms, attributes and their values, as well as the required specification level. Thus, MST portrays its structural information in a visible manner that is likely to promote uniform interpretation by expert users. This explicit and implied knowledge is just as important as the collection of terms in the MST. This information is essential for applications to be able to make full use of the terminology.

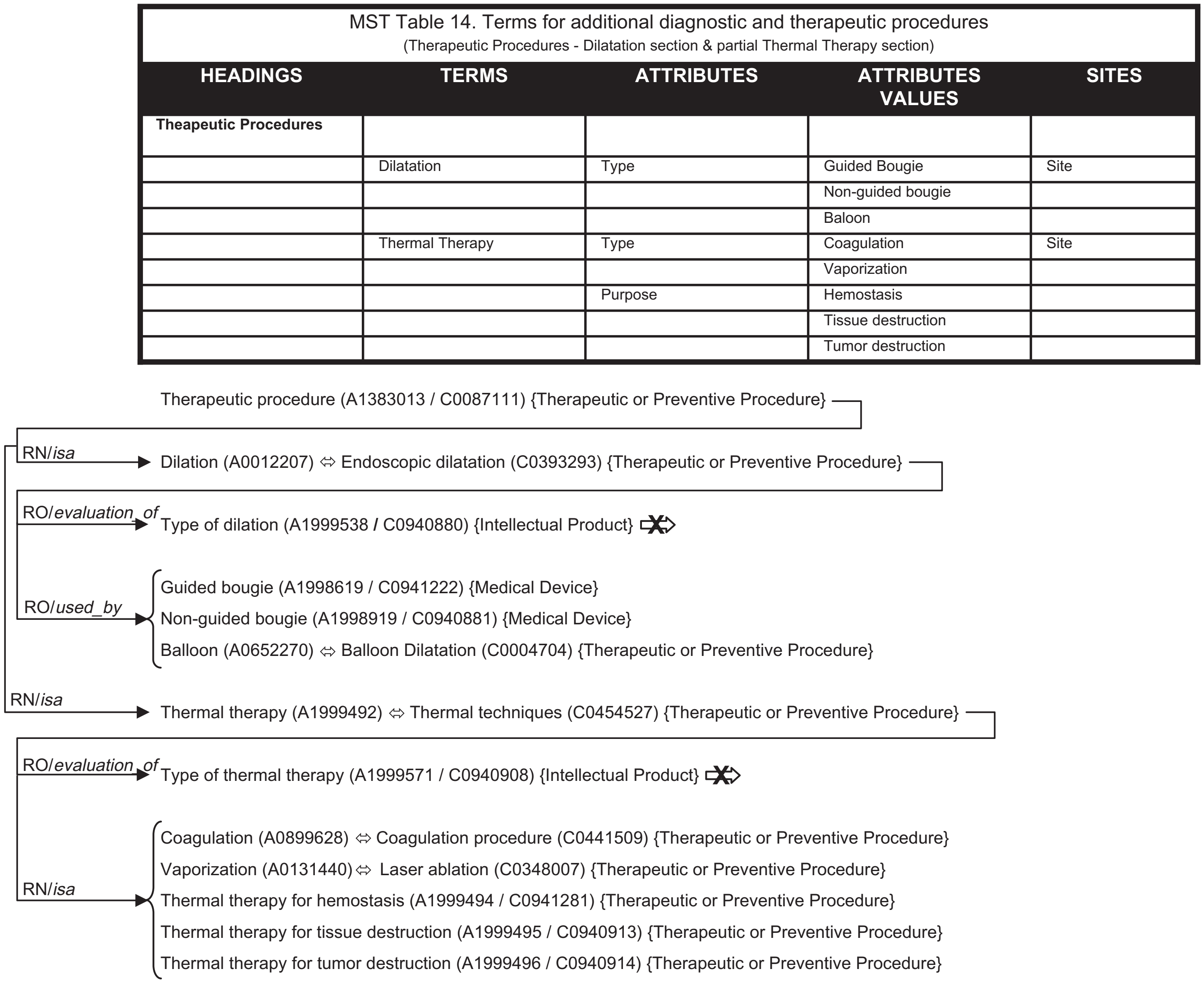
As demonstrated by our findings, while the UMLS MST integration retained all MST terms, the implementation does not allow for accurate recreation of the MST structure and the inter-relationships between its terms. Thus, while the MST was created to promote standardized knowledge capturing and sharing, the UMLS MST integration falls short in structure and content and cannot be truly used by applications as the MST was intended. Therefore, developers cannot rely on MTHMST to be an authentic representation of the MST.

Many UMLS source vocabularies are available in a format that may not be immediately reusable for a developer. They may be provided in a paper-based format or some non-concept oriented electronic format. Both may require manipulation to be used. This manipulation provides an opportunity for further introduction of "noise" since each developer may process them differently. The UMLS is uniquely positioned so that its terminology integration process can offer a centralized, electronically accessible model of such terminologies. However, in order to position itself as such, it has to maintain a close compatibility with the source.

Retaining the authentic modeling of the source MST in an application accessible environment will clearly benefit all applications that desire to take advantage of the knowledge incorporated in it. However, the current modeling of MST in the UMLS is cumbersome and non-compliant with the source MST. Thus, MTHMST is not as valuable for GIE application developers. Since the integration process of source vocabularies into the UMLS has not been very transparent, this case study of the MST along with suboptimal modeling of LOINC [14] may indicate the need to examine other source vocabularies, along with providing better transparency of the integration process. If the UMLS is to integrate source terminologies in a manner that will serve for more than concept mapping, more attention should be applied to maintain the original modeling and intent of the source and to perform all integrations in a consistent manner.

5.2. Singletons and modeling discrepancies

Most of the imported MST terms (80%) were new UMLS concepts. We identified 210 UMLS singletons and 286 MST singleton terms which lack hierarchical relationships (PAR, CHD, RB, RN) within the UMLS or the MST, respectively. These represent more than 10% of the MST terms. Additionally, 65 of the MST singleton terms were found to be super-singletons. Super-singletons lack any MST-derived hierarchical or lateral relationships. Closer examination of the MST shows that at least between table titles and table headings, hierarchical relationships are common, and at times additional hierarchical levels can be found within the same tables.



MST Table 14. Terms for additional diagnostic and therapeutic procedures
(Therapeutic Procedures - Dilatation section & partial Thermal Therapy section)

HEADINGS	TERMS	ATTRIBUTES	ATTRIBUTES VALUES	SITES
Theapeutic Procedures				
	Dilatation	Type	Guided Bougie	Site
			Non-guided bougie	
			Baloon	
	Thermal Therapy	Type	Coagulation	Site
			Vaporization	
		Purpose	Hemostasis	
			Tissue destruction	
			Tumor destruction	

Therapeutic procedure (A1383013 / C0087111) {Therapeutic or Preventive Procedure}
RN/isa
Dilation (A0012207) ⇔ Endoscopic dilatation (C0393293) {Therapeutic or Preventive Procedure}
RO/evaluation_of
Type of dilation (A1999538 / C0940880) {Intellectual Product} ⇏
RO/used_by
Guided bougie (A1998619 / C0941222) {Medical Device}
Non-guided bougie (A1998919 / C0940881) {Medical Device}
Balloon (A0652270) ⇔ Balloon Dilatation (C0004704) {Therapeutic or Preventive Procedure}
RN/isa
Thermal therapy (A1999492) ⇔ Thermal techniques (C0454527) {Therapeutic or Preventive Procedure}
RO/evaluation_of
Type of thermal therapy (A1999571 / C0940908) {Intellectual Product} ⇏
RN/isa
Coagulation (A0899628) ⇔ Coagulation procedure (C0441509) {Therapeutic or Preventive Procedure}
Vaporization (A0131440) ⇔ Laser ablation (C0348007) {Therapeutic or Preventive Procedure}
Thermal therapy for hemostasis (A1999494 / C0941281) {Therapeutic or Preventive Procedure}
Thermal therapy for tissue destruction (A1999495 / C0940913) {Therapeutic or Preventive Procedure}
Thermal therapy for tumor destruction (A1999496 / C0940914) {Therapeutic or Preventive Procedure}

Fig. 5. On top is part of the Therapeutic Procedures section of MST Table 14. The bottom section represents an effort to recreate the content of that section utilizing available MTHMST hierarchical and other relationships. Relationship type and attribute are marked above the arrows. Terms and concepts are presented separately if different than the preferred name (Term ID) ⟺ (Concept ID) {semantic type}. ⇏ indicates concepts that lack relationships to their respective diseases and attributes column content.

Therefore, it is surprising that a relatively large segment of MTHMST lacks any such hierarchical associations, and may indicate non-optimal modeling during the MST integration process. Many singletons seem to have been removed from their MST context, and we speculate that this may have contributed to their singleton status since their inter- and intra-MST table relationships were lost in the process. For example, in MST Table 18, acute pancreatitis is one of the diseases that can be used to specify reasons to perform ERCP. It should be specified with an attribute such as ‘‘suspected” or ‘‘exclusion of”. However, **Acute pancreatitis** (A0242598), one of the MST super-singletons, lacks any relationships that would convey that information, and is mapped out of context.

5.2.1. Fully specified terms

A cardinal feature of the UMLS integration of the MST (as well as many other source terminologies) was the desire to create fully specified terms for every combination of headings, terms, attributes and attribute values according to the MST columns of specific tables. This was based on the assumption that this is required so that each term will be fully understood by human or machine agents [11]. However, a fully specified term is still an alphanumeric string value, and while it can be intuitively understood by human agents or may be suitable for natural language processing, it may not be the optimal substrate for other types of applications.

The creation of fully specified terms for the MST UMLS integration resulted in a side effect. In the process, actual attributes, attribute values and modifiers lose their identity as such and become terms. For example, MST Table 12 (Table 1) lists terms of findings for the pancreas. The attribute value **Multiple** for Abnormalities|Cavity|Number|Multiple has been converted into the fully specified **Multiple pancreatic cavities** [A1998871]. The UMLS MST integration attempted to retain the original intent by creating MST Classes (MSC) assigned to each MST term. Thus, **Multiple pancreatic cavities** [A1998871] has an assigned MSC of {Finding.Attribute.Value}. However, **Multiple pancreatic cavities** is not the attribute value **Multiple** and any application that will attempt to utilize MTHMST to recreate the list of allowable attribute values for the finding of **Pancreatic cavity** will face difficulty comprehending which component of the string **Multiple pancreatic cavities** is the actual value. This represents a realistic scenario if developers will attempt to use the UMLS MST component to create structured data entry screens for GIE examinations within applications.

5.2.2. Reusability

The UMLS MST integration put much effort to make explicit the implied relationships between MST terms and across MST tables. However, as our findings demonstrate, this conversion is not transparent or intuitively understood, nor does it allow recreation of accurate and complete MTHMST data sets and structures that are fully compatible with MST sets. Since the schema of the MST is not captured and conceptualized in MTHMST, intelligent agents cannot fully access and utilize MST knowledge in the UMLS. Programmers wishing to develop applications using MTHMST will be required to have extensive intrinsic knowledge of the original MST and will have to program many tweaks and adjustments into their logic to make the most out of the current implementation.

5.3. Approach for integration

We would like to propose two main reasons for the problems encountered with MTHMST. First, as mentioned above, while the MST is in effect a concept oriented terminology, this information is embedded and implied in the collection of its table structure. Despite all the different measures by Tringali et al. [11], there seems to have been no explicit stage at which the MST was formalized as concept oriented terminology before its integration into the UMLS MTH. The creation of fully specified terms, extraction of relationships, and integration with existing META concepts, was done in one continuous move. We believe that formalizing the conceptual structure is a required and crucial distinct first step, since it crystallizes the structure of the hierarchical tree and the lateral relationships between its concepts. While most UMLS source terminologies are already formalized as concept oriented terminologies at the time of integration, this clarity was missing at the time of the MST integration. We hypothesize that a two step approach, in which every source terminology first exists as an explicitly formulated concept oriented terminology will result in better quality, second step integration.

5.4. The significance of fully specified terms

We would like to propose that not every item of the source terminology must be integrated as a fully specified term as was accomplished with MTHMST. Terminology items such as term modifiers, whether they are defined as modifiers, attributes or attribute values as in the case of the MST do not necessarily have to be incorporated as the fully specified combinations of each allowable permutation with the terms to which they are assigned. No additional value is captured by adding **Multiple pancreatic cavities** and **Single pancreatic cavity** as independent concepts. The same information can be conveyed by assigning a relationship type and a relationship attribute that describe the nature and the meaning of the association related to the quantitative value of the finding. This follows the exact modeling structure of the MST (see Fig. 4). Mimicking the UMLS, SNOMED CT [23] has extensive capabilities to accommodate such modeling. SNOMED is based on description logic (DL) and its qualifiers are organized in a separate hierarchy, intended for use through lateral relationships and post-coordination. However, the vast majority of SNOMED's fully defined quantitative concepts are similarly pre-coordinated and not conceptualized accordingly.

Examination of MST Table 12 reveals that the table name and first and second column values actually create a hierarchical structure, while column 3 (Attributes) defines various types of relationships, column 4 (Attributes values) lists the modifiers that apply for each such relationship and column 5 defines a site relationship between column 2 terms and terms that are listed in the pancreas section of MST Table 3 – Sites for location of findings on ERCP (implied). The UMLS framework is ready to accommodate such modeling. As a matter of fact, many of the MST modifiers already exist in the UMLS as concepts. Thus, existing modifiers can be reused in multiple associations, significantly reducing the number of new fully specified terms required to be created. This would have a beneficial effect in terms of maintenance and performance if similarly applied to other source terminologies.

The MST, in some respects, is not unlike LOINC. Both are compositional systems with the MST placing more restrictions on possible combinations than LOINC. As Bodenreider [14] summarizes, the integration of LOINC in the UMLS remains suboptimal, but the naming convention adopted for the LOINC UMLS integration retains the original compositional nature of its names whereas the MST effort puts significant emphasis on converting MST terms to fully specified names. While it can be argued that compositional names such as in LOINC are not optimal for natural language processing, there are other means for controlled terminologies to support applications, such as the network of relationships between concepts. We propose that retaining complete concept definitions, including the exact semantic types of various concept attributes, is essential and must be preserved if integrated terminologies are to retain their usefulness as in their independent state.

5.5. Retaining original functionality

The MST can be defined as an interface terminology [24–27]. Such terminologies must balance between pre- and post-coordination [27], and encapsulate many functional and clinical aspects intrinsic to the domains they serve. It has been demonstrated [25,26,28] that once information is captured by an interface terminology, it can than be mapped to reference terminologies [28]. However, concept mapping is quite a different task than maintaining the original functionality of the interface terminology within the reference terminology framework, as attempted by the MST integration in the UMLS [11].

Had the UMLS sole goal been simply to provide term mapping, very minor issues would have arisen based on the findings above. However, the UMLS goal is also to provide cross-functionality between terminologies, for which the precise representation of inter-concept relationships is of the utmost importance. The MST is designed to support annotation of gastrointestinal examination reports in a manner that provides a consistent structure and content across the various applications that choose to utilize it. The MTHMST integration stated that it aims to provide "a new (but MST-compatible) machine-readable terminological tool" [11]. However, if compatibility between the source terminology and its internal UMLS representation is not maintained, the usefulness of the integration for clinical and other healthcare applications might be called into question.

5.6. Transparency and authenticity assessment

Much research has been conducted regarding the configuration of META concepts and terms. Lexical, semantic and structural methods have been applied to detect potential errors in UMLS concepts. However, this body of knowledge looks at META concepts as independent concepts, out of their source's context. While the UMLS currently incorporates more than 150 source vocabularies, very little has been formally published regarding the actual integration process and the accommodations made during such a process. Moreover, no research directly addresses the issue of source authenticity. As the example in Fig. 5 clearly demonstrates, hierarchical and lateral relationship types and their respective attributes should serve as excellent compatibility indicators since they are directly based on same or similar relationship qualities in source terminologies.

5.7. MST – past and present

The MST has progressed over the years. OMED currently offers MST version 3.0 [29,30]. It is structured and published in a very similar manner to MST version 2. However, the UMLS integration of MST is frozen in time, and has not been updated since 2002. Moreover, a PubMed search did not yield any published results that indicate that MTHMST has been utilized for any project or application. At the same time, various gastroenterology systems and groups utilize MST, and MST-based ability to annotate GIE reports remains desirable [31,32].

MST curators were not closely involved with its integration into the UMLS. We speculate that a more collaborative approach would have resulted in a more authentic and viable product.

6. Conclusions

The integration process of the MST into the UMLS included certain adaptations that resulted in an MTHMST that is not optimally authentic compared to the source MST. In turn, this is likely to reduce the usability of the MTHMST for application developers who might want to benefit from the significant knowledge implied, but not explicitly represented in it. For reusable and authentic representation of source terminologies we propose that retaining complete concept definitions, including the exact semantic types of various concept attributes, is essential and must be preserved, even at the cost of the creation of less than fully specified terms in the UMLS. Enhanced transparency of the source vocabulary integration process will enable UMLS users to better grasp potential differences between the UMLS representation and the original source modeling.

Acknowledgment

This work was partially supported by the United States National Library of Medicine under grant R 01 LM008445-01A2.

References

[1] Unified Medical Language System – fact sheet. Available from: http://www.nlm.nih.gov/pubs/factsheets/umls.html.
[2] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993;32(4):281–91.
[3] McCray AT, Miller RA. Making the conceptual connections: the Unified Medical Language System (UMLS) after a decade of research and development. J Am Med Inform Assoc 1998;5(1):129–30.
[4] Lindberg C. The Unified Medical Language System (UMLS) of the National Library of Medicine. J Am Med Rec Assoc 1990;61(5):40–2.
[5] About the UMLS resources. Available from: http://www.nlm.nih.gov/research/umls/about_umls.html.
[6] Nelson SJ, Tuttle MS, Cole WG, Sherertz DD, Sperzel WD, Erlbaum MS, et al. From meaning to term: semantic locality in the UMLS Metathesaurus. Proc Annu Symp Comput Appl Med Care 1991:209–13.
[7] Tuttle MS, Sherertz DD, Erlbaum MS, Sperzel WD, Fuller LF, Olson NE, et al. Adding your terms and relationships to the UMLS Metathesaurus. Proc Annu Symp Comput Appl Med Care 1991:219–23.
[8] McCray AT, Aronson AR, Browne AC, Rindflesch TC, Razi A, Srinivasan S. UMLS knowledge for biomedical language processing. Bull Med Libr Assoc 1993;81(2):184–94.
[9] McCray AT, Nelson SJ. The representation of meaning in the UMLS. Methods Inf Med 1995;34(1–2):193–201.
[10] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Stud Health Technol Inform 2001;84(Pt. 1):216–20.
[11] Tringali M, Hole WT, Srinivasan S. Integration of a standard gastrointestinal endoscopy terminology in the UMLS Metathesaurus. In: Proceedings of the AMIA symposium; 2002. p. 801–5.
[12] Fung KW, Hole WT, Nelson SJ, Srinivasan S, Powell T, Roth L. Integrating SNOMED CT into the UMLS: an exploration of different views of synonymy and quality of editing. J Am Med Inform Assoc 2005;12(4):486–94.
[13] Zhang S, Bodenreider O. Experience in aligning anatomical ontologies. Int J Semant Web Inf Syst 2007;3(2):1–26.
[14] Bodenreider O. Issues in mapping LOINC laboratory tests to SNOMED CT. In: Proceedings of the AMIA annual symposium; 2008. p. 51–5.
[15] Lomax J, McCray AT. Mapping the gene ontology into the unified medical language system. Comp Funct Genomics 2004;5(4):354–61.
[16] Delvaux M. The minimal standard terminology for digestive endoscopy: introduction to structured reporting. In: Hagenmüller F, Manns MP, Musmann HG, editors. Medical imaging in gastroenterology and hepatology. Springer; 2002. p. 112–24.
[17] Maratka Z. The OMED data base: standard for nomenclature. Endoscopy 1992;24(Suppl. 2):455–6.
[18] Maratka Z. Terminology, definitions and diagnostic criteria in digestive endoscopy. With the collaboration of the members of the Terminology Committee of the World Society of Digestive Endoscopy/OMED. Scand J Gastroenterol Suppl 1984;103:1–74.
[19] Crespi M, Delvaux M, Schapiro M, Venables C, Zwiebel F. Minimal standards for a computerized endoscopic database. Am J Gastroenterol 1994;89:144–53.
[20] Delvaux M, Korman LY, Armengol-Miro JR, Crespi M, Cass O, Hagenmüller F, et al. The minimal standard terminology for digestive endoscopy: introduction to structured reporting. Int J Med Inform 1998;48(1–3):217–25.
[21] UMLS reference manual. Available from: www.nlm.nih.gov/research/umls/meta2.html.
[22] Bodenreider O. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. In: Bakken S, editor. Proceedings of the 2001 AMIA annual symposium, 2001 November, Washington, DC; 2001. p. 57–61.
[23] SNOMED CT. Available from: http://www.ihtsdo.org/snomed-ct/.
[24] McDonald FS, Chute CG, Ogren PV, Wahner-Roedler D, Elkin PL. A large-scale evaluation of terminology integration characteristics. In: Proceedings of the AMIA symposium; 1999. p. 864–7.
[25] Chute CG, Elkin PL, Sherertz DD, Tuttle MS. Desiderata for a clinical terminology server. In: Proceedings of the AMIA symposium; 1999. p. 42–6.
[26] Campbell KE, Das AK, Musen MA. A logical foundation for representation of clinical data. J Am Med Inform Assoc 1994;1:218–32.
[27] Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc 2006;13(3):277–88.
[28] Spackman KA, Campbell KE, Cote RA. SNOMED RT: a reference terminology for health care. In: Proceedings of the AMIA annual fall symposium; 1997. p. 640–4.
[29] MST 3.0. Available from: http://omed.org/index.php/resources/re_mst/.
[30] Aabakken L, Rembacken B, LeMoine O, Kuznetsov K, Rey JF, Rösch T, et al. Minimal standard terminology for gastrointestinal endoscopy – MST 3.0. Endoscopy 2009;41:727–8.
[31] Groenen MJ, Kuipers EJ, van Berge Henegouwen GP, Fockens P, Ouwendijk RJ. Computerisation of endoscopy reports using standard reports and text blocks. Neth J Med 2006;64(3):78–83.
[32] Liu D, Cao Y, Kim KH, Stanek S, Doungratanaex-Chai B, Lin K, et al. Arthemis: annotation software in an integrated capturing and analysis system for colonoscopy. Comput Methods Programs Biomed 2007;88(2):152–63.