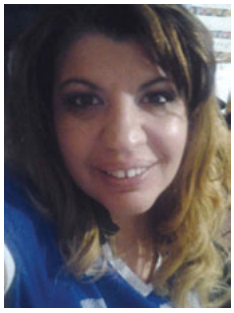
Interoperability and Mapping Between Knowledge Organization Systems: Metathesaurus—Unified Medical Language System of the National Library of Medicine

Julietti de Andrade* and Marilda Lopes Ginez de Lara**

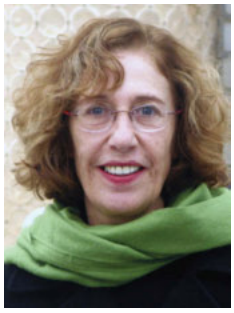
* University of São Paulo, Institute of Orthopaedic Surgery, Hospital das Clínicas, Faculty of Medicine, Rua Ovídio Pires de Campos 333, São Paulo, Cerqueira Cesar, Brazil, <julietti.andrade@gmail.com>

** University of São Paulo, School of Communications and Arts, Postgraduate Program in Information Science, Cidade Universitária 443, São Paulo, CEP, Brazil 05508-020 <larama@usp.br>

Julietti de Andrade holds a PhD in information science from the Arts Communication School of University of São Paulo-ECA-USP with Sandwich doctorate at the Faculty of Translation and Documentation, University of Salamanca, Spain. She is the Director of Library Service and Scientific and Educational Documentation Institute of Orthopaedic Surgery, Hospital das Clínicas, Faculty of Medicine, University of São Paulo. She works in the planning and management unit and teaches and conducts research in knowledge organization, information retrieval and health information sources.



Marilda Lopes Ginez de Lara holds a PhD in communication sciences from the School of Communications and Arts, University of São Paulo. She is an associate professor at the postgraduate program in information science, School of Communications and Arts, University of São Paulo. Her teaching and research area is cultural references of knowledge organization systems.



Andrade, Julietti de and Marilda Lopes Ginez de Lara. 2016. "Interoperability and Mapping Between Knowledge Organization Systems: Metathesaurus—Unified Medical Language System of the National Library of Medicine." *Knowledge Organization* 43, no. 2: 107-112. 16 references.

Abstract: This paper is aimed at assessing the potential of interoperable knowledge organization systems to respond to search strategies in order to retrieve information from databases in the areas of health and biomedicine. An analysis was done on the semantic consistency of synonym grouping of a term selected from the Metathesaurus, the Unified Medical Language System of the National Library of Medicine, based on the characteristics of equivalence proposed in ISO 25964: 2: 2011 and based on the following categories: semantic, morphological, syntactic and typographical variations. This paper highlights the importance of understanding the results of automatic mapping as well as the need for characterization, evaluation and selection of equivalences for preparation of consistent search strategies and presentation of search results in scientific work methodologies.

Received: 9 December 2015; Revised 12 December 2015; Accepted 19 December 2015

Keywords: terms, search, knee arthroplasty, Metathesaurus, synonyms, term equivalences

1.0 Introduction

One of the difficulties in carrying out bibliographical research aimed at retrieving information to produce scientific knowledge is mapping and selection of appropriate terms and concepts to create search strategies. Within these activities, there is the complexity of the language—the raw material of knowledge organization systems (KOSs) —which is represented in morphological variations (grammatical classes), semantics (limits of meaning), syntax (position of terms within the descriptor, use of punctuation), and typography (Andrade 2015). These

variations exist in terms that, aligned with the obstacles of retrieving information based on character matching result in disjoining of documents of the same semantic nature, might result an inconsistent sample of documents retrieved from the scope of a particular database.

Considering this problem, this paper is aimed at assessing the potential of interoperable knowledge organization systems (KOSs), at responding to search strategies to retrieve information from databases specialized in the areas of health and biomedicine. We selected the U.S. National Library of Medicine (NLM) Metathesaurus, which contains 151 KOSs, making it a concrete example of mapping and promotion of interoperability, the sort of requirements needed for a semantic web to function. To perform the evaluation, we tested the search results of the aforementioned instrument based on an analysis of a group of equivalencies for a selected term.

KOSs, especially thesauri and lists of subject headings, are usually used in two ways in databases:

1. KOS not integrated into the database:
Used in inputting into the system, when indexing the document, by entering select terms in a specific field, such as field 650 in the MARC21 (Machine Readable Cataloging) format; in output from the system when specialists and librarians use the KOS to consult terms considered appropriate for recovery of documents and information on certain subjects.

This form of using the KOS creates two situations. First, there must be no errors in typing and registering a term with distinct typographical forms (e.g., uppercase, lowercase), since each manner of registration creates an input and, therefore, a mismatch in retrieving documents on the same subject. In this context, Lopes (2002, 61) cites a study by Bourne (1977) on the impact of spelling errors in online searches, where the author states that "the number of spelling errors, variant forms of words and even spelling errors for indexing terms considerably affect search results." Second, all equivalent terms considered appropriate from the semantic, syntactic, morphologic and typographic standpoints must be used in building search strategies so that a consistent result can be obtained.

2. In the case of retrieval by hierarchies, some databases allow for restriction on the retrieval of the terms desired and not the terms subordinated to it (if any), as is the case with PubMed (U.S. NLM 2013a) and Embase (Elsevier 2013).

Nevertheless, even when the database offers the chance to do searches using KOSs, usually only the KOS in the database is used. For instance, the PubMed database allows for searches to be done using *Medical Subject Headings* (*MeSH*), which currently contains 27,455 descriptors (U.S. NLM 2015) and Embase contains the Emtree thesaurus, which has around 60,000 terms (Elsevier 2013).

In this sense, one of the ways performance can be enhanced in the search and information retrieval processes is by developing and implementing interoperability and mapping among KOSs so that the sample of terms to be used in search strategies can be expanded and substantiated through simultaneous use of KOSs that already exist and through both databases containing KOSs for the search and databases that only retrieve based on character matches.

2.0 Interoperability and mapping between knowledge organization systems

The main standards for building a KOS define interoperability as the ability of two or more systems to exchange and use information exchanged without a special effort by any of the systems. The main standards referred to here are Association for Library Collections and Technical Services 2000; National Information Standards Organization 2005; BSI Group 2007; International Standards Organization (ISO) 2011a. The ISO 25964-2:2011 standard highlights the objective of interoperability in information retrieval as being the possibility of using an expression formulated in a KOS converted into a corresponding expression in one or more KOSs, while also underscoring the establishment of equivalences for the success of the operation.

Mappings between the KOSs, such as the establishment of relationships between them, are covered by the ISO 25964-2:2011 standard. Among thesauri in particular, the standard identifies three main types of mapping (ISO 2011b): these are equivalence, hierarchical and associative, with equivalence being the most common and necessary type. It recommends that equivalences between thesauri should be established when the corresponding concepts are found in two or more different KOSs, with no difference in status between the concepts or between the terms preferred or notations that represent them.

In order to contribute to identification of concepts from the semantic standpoint, the ISO 25964-2:2011 standard organizes equivalences into types and levels. The types correspond to mapping of simple and compound equivalences, with there being recommendations for compound equivalence regarding establishment of intersection for compound equivalences, cumulative compound equivalence and compound equivalence involving target vocabulary. Levels include mapping of exact, inexact and partial equivalences. Exact equivalence occurs "when concepts can be used interchangeably through all the applications

that can be envisaged for the mapping;" inexact equivalence is when two concepts corresponding in two or more vocabularies are not exactly the same; and partial equivalence is related to generic or specific characteristics of the meanings of the terms.

3.0 Unified Medical Language System-UMLS: Metathesaurus

Unified Medical Language System (UMLS) is a product of the U.S. NLM that is made up of a set of files and software that gather vocabulary in the areas of health and biomedicine, as well as standards to allow for interoperability between computer systems. Databases and associated software tools that are part of the UMLS are aimed primarily at the developers in the process of building or improving systems whose are: creation, processing, retrieval and integration of information and data in the health area. Associated tools help developers to use and customize UMLS databases in private applications. Use of the UMLS requires that a license be requested from the NLM, which establishes all of the ways to use this tool. Only American citizens can download the KOSs that are part of the UMLS (U. S. NLM 2009, 2013bc, 2014), with citizens of other countries being allowed to check the databases made available through a valid license granted for one year.

The Metathesaurus is one of the components of the UMLS (Martínez Tamayo et al. 2011), a project developed by the U.S. NLM in 1986 and continually updated. It is characterized as a database that currently contains 151 multilingual KOS, integrating concepts and information on them in the areas of health and biomedicine. It presents a set of files organized by concept or meaning that relates alternative names and visions for the same concept from different KOS sources. Designed to be used in creating applications related to automatic indexing, the Metathesaurus (U. S. NLM 2009) is built using various electronic versions of thesauri, classifications, code sets and lists of controlled terms used in patient care, health services billing, public health statistics, indexing and cataloguing of biomedical literature, and basic health and clinical research services.

In addition to the Metathesaurus, the system has two other components: the Semantic Network and the SPECIALIST Lexicon with Lexical Tools. The Semantic Network provides a categorization of concepts and defines the set of relationships between Metathesaurus semantic types. At least one semantic type is attributed to each concept. The semantic types include anatomical structure, biologic and chemical functions, illness or syndrome, labs or exam results, medical devices and organisms. The SPECIALIST Lexicon provides lexical information for the Natural Language Processing System (NLP), whose function is to mediate between the language used by users and the language in the sources of biomedical information. Although it is a general language lexicon, it includes biomedical terms. The SPECIALIST Lexicon input for each word or term registers the syntactic, morphological and orthographical information needed by the SPECIALIST NLP System.

The SPECIALIST Lexicon operates along with the lexical tools, which are designed to address the high degree of variability in natural language words and terms (U. S. National Library of Medicine 2009):

> The Lexical Tools are designed to address the high degree of variability in natural language words and terms. Words often have several inflected forms which would properly be considered instances of the same word. The verb "treat," for example, has three inflectional variants: "treats" the third person singular present tense form, "treated" the past and past participle form, and "treating" the present participle form. Multi-word terms in the Metathesaurus and other controlled vocabularies may have word order variants in addition to their inflectional and alphabetic case variants. The Lexical Tools allow the user to abstract away from this sort of variation.

3.1 Types of Metathesaurus search and retrieval

Metathesaurus search can be done by term, by concept unique identifier (CUI) and by code corresponding to numbering related to atoms, which can be equivalences or related terms. It is also possible to select the version (release) and the KOS (source) of the desired Metathesaurus. KOSs are represented by acronyms listed alphabetically. A term search can be done by word, by approximate match, by exact match, by normalized string, by normalized word, by truncation right and by truncation left.

The interface of the results is organized into three categories: basic view, report view and raw view. The basic view and report view categories show common metadata in the search result, such as concept, semantic type and definition. The difference between the two is that in basic view there is a list of synonyms and a list of relations to the term selected for consultation; while in the report view, atoms are listed instead of synonyms, in addition to two other lists, entitled contexts and contextual relationships. Note that the Metathesaurus considers a series of occurrences that integrate semantic, syntactic, morphologic and typographic characteristics as synonyms, which include alphabetical and typographic ordering (terms organized in the sequence of upper case, upper and lower case and lower case), and not just semantic

aspects as described in the main standards for KOS construction. These variations correspond to how the terms are registered in their respective KOS sources.

3.2 Assessment of Metathesaurus equivalences

Assessment of the Metathesaurus equivalences was preceded by an analysis of the purposes and characteristics of the instrument to be analyzed, in relation to the product objectives, target public and structure of the KOS that integrate it, as well as identification of the search types and retrieval methods through execution of open searches with specialty terms from orthopedics and traumatology, more specifically those regarding knee arthroplasty.

Results of open searches represented by the large numbers of terms retrieved in the search types geared towards retrieval based on character match, the irrelevant results with and without KOS selection in relation to the objective, which was to exclude equivalences from languages other than English from the set of retrieved synonyms of the selected term, and the content of the result interfaces (synonyms and atoms) were the foundations for establishing the search and retrieval criteria, which were:

- Criteria for KOS inclusion and exclusion, search types and result interfaces, as well as definition of the terms used for the search;
- Selection of the search term knee arthroplasty and its respective equivalent terms from the Emtree thesaurus in the Embase Biomedical Database Elsevier (2013) thesaurus, because this thesaurus is not part of the Metathesaurus. A choice was made to work with KOS terms that are not part of the Metathesaurus in order to verify how much automatic mapping is able to integrate possible equivalences;
- Establishment of criteria for analysis of results;
- Application of search strategies along with records and systemization of results;
- Establishment of a record-keeping model and systemization of search results;
- Assessment of search results based on use of the ISO 25964:2:2011 standard and on semantic, morphologic, syntactic and typographic aspects to identify and select equivalent terms resulting from mapping between the KOSs of the products analyzed;
- Re-elaboration of systematization of results when necessary; and,
- Inclusion and exclusion of terms.

After carrying out the search test and identifying results, the "Knee Replacement Arthroplasty (Procedure)" term was chosen along with the respective 146 synonyms, which were classified as exact, inexact, partial and/or related terms.

In the first stage, synonyms were analyzed in relation to their respective languages, which resulted in inclusion of sixty-three terms and exclusion of eighty-three terms. Based on these results, two tables were made: included synonyms and excluded synonyms.

In the second stage, the included synonyms table was organized by using the sequential number of synonyms for the term chosen for analysis. Next, a check was done for duplication by selecting and pasting the synonym term from the Metathesaurus search result interface and verifying the list registered in Word. This procedure was necessary because of the presentation method used by the database which, when using typographic ordering, ends up creating duplication and some physical distance between terms that are the same but registered with different letter types.

In the third stage, synonyms were organized into categories included and excluded. Those included were classified by levels of equivalence: exact, inexact and partial according to the ISO 25964:2:2011 standard, as initially explained. An individual analysis of each synonym allowed terms to be identified that, although considered equivalent in the mapping, are actually related terms.

After analysis and individual classification of each synonym by level of equivalence and/or related term, the terms were regrouped with the aim of identifying the levels of equivalences by group of similar or same synonyms from the typographical, morphological or syntactic standpoint, for later inclusion or exclusion. Fourteen groups were created. Sequential numbering established in creation of the table of included terms was maintained when organizing the tables related to synonym groups in order to visualize duplications and record the physical distance between forms of the same term, which could be greater if using the numbering of the general list of synonyms, which also contains excluded terms. This reorganization allowed similarities and differences to be found in the synonym groups and the resulting exclusion of related terms and duplications. Table 1 shows an example of a synonym group included and excluded from the "Knee Replacement Arthroplasty (Procedure)" term from the Metathesaurus (UMLS) corresponding to the term "Arthroplasty knee."

After reorganization of the terms and systematization of classifications, of the total of 63 synonyms included, twelve terms were excluded: ten because they were duplications caused by typographical variation and two that were related terms. Therefore, of the initial list of one hundred forty-six synonyms, four were considered exact equivalences, forty-two partial equivalences in relation to

Synonyms of Knee Replacement Arthroplasty (procedure)	Criteria for Inclusion and Exclusion
7. Arthroplasty knee	Included – Exact equivalence in relation to Knee arthroplasty. – Partial equivalence for Knee Replacement Arthroplasty (procedure). – Semantic Variation: substitution not emphasized. – Syntactic variation (position of the words within the term) – Morphological variation (with and without preposition)
8. Arthroplasty of knee	
9. Arthroplasty of knee (procedure)	
11. Arthroplasty of the Knee	
15. KNEE ARTHROPLASTY	
52. knee arthroplasty (treatment)	
10.Arthroplasty of knee	Excluded – Duplications
12.Arthroplasty of the knee	
26.Knee arthroplasty	
49.arthroplasty of knee	
50.arthroplasty of the knee	
51.knee arthroplasty	

Table 1. Synonyms of "Knee Replacement Arthroplasty (Procedure)" and criteria for inclusion and exclusion.

the terms used in the "Knee Arthroplasty" search and "Knee Replacement Arthroplasty (Procedure)" retrieval and five inexact equivalences, totaling fifty-one terms considered equivalent for construction and application of sensitive search strategies in health databases that operate with the English language.

Note that terms were included which, when analyzed in the context of the group, could be considered the same from a semantic standpoint, but morphologically, syntactically and typographically different. Non-inclusion of these terms in a search strategy creates differences in results, especially in databases that operate with automatic indexing and retrieval based on character match.

In relation to verification of equivalent terms from the Emtree thesaurus for "Knee Arthroplasty" in the Metathesaurus, findings showed that only four of the set of ten Emtree terms (see Table 2), including the preferred term "Knee Arthroplasty," were contained in the set of one hundred forty-six synonyms, with two ("Arthroplasty, Knee" and "Knee Arthroplasties") which can be considered as exact equivalences from the semantic standpoint with morphological variation (number) and syntactic variation (with comma) in relation to the terms used in search and retrieval. Put another way, even with the mapping of synonyms, there are still terms left out that could be considered equivalent. This is due to the many possibilities of syntactic and morphological variation of a term, especially in the English language. The same occurs when analyzing the terms "Knee Joint Replacement" and "Knee Joint Replacements," which could also be mapped as synonyms of "Knee Replacement Arthroplasty (Procedure)," since "Knee Joint Replacement" operation was mapped as a synonym. The term "Knee Joint Replacement" was recovered in the Metathesaurus without definitions, but the analysis of relations in the Basic View retrieval interface allows for this conclusion.

Emtree preferred term and equivalents	Is it listed as a synonym of Knee Replacement Arthroplasty (Procedure) in the Metathesaurus?
arthroplasty, replacement, knee knee arthroplasty knee replacement knee replacements	Yes
arthroplasty, knee knee arthroplasties knee joint replacement knee joint replacements knee reconstruction Reconstruction, knee	No

Table 2. Comparison between equivalent terms in Emtree and Metathesaurus.

The terms "Knee Reconstruction" and "Reconstruction, Knee," treated as equivalents of "Knee Arthroplasty" in the Emtree, referred to concepts of "Anterior and Posterior Cruciate Ligament Reconstruction" in the Metathesaurus and not specifically to "Knee Arthroplasty" which, according to the definitions analyzed during this research, are focused on replacement and not on reconstruction, justifying the fact that they are not part of the group of term synonyms analyzed.

4.0 Conclusions

Note that even with multiple mappings done in an effort to group the possible semantic, syntactic, morphological and typographic variations of the terms, there were still some equivalences that were not mapped.

Organization of the synonyms in alphabetical order followed by typographical order (upper case, upper and

lower case and lower case) creates a list with syntactic (punctuation and position of words in the term) and morphologic (number and use of prepositions) variations. One of the commitments of the Metathesaurus is to maintain the terms, whether they are preferred, equivalent or related, in the original form of the KOS source, which causes the typographical variations found and, consequently, duplications.

Use of interoperated and mapped KOSs to identify and select equivalent terms to create search strategies in health is considered relevant, from the point of view of document retrieval as well as to present search results when composing methodology for scientific papers. A sample of well-delimited terms certainly offers more quality to retrieval and presentation of data.

The conclusion was reached that the Metathesaurus fulfills the interoperability and mapping function among the KOSs by allowing for simultaneous search of terms in various KOSs, categorizing and establishing hierarchical, equivalence and associative relationships among all terms from the KOS sources in the database, therefore justifying its name. Nevertheless, results of automatic mapping, information retrieval methods based on character match and fixed organization of metadata in the results interface require that the results be assessed, especially insofar as the terms and equivalences retrieved from the semantic, morphological, syntactic and typographical standpoints are concerned.

Understanding of the precepts of the semantic web and of interoperability, as well as analysis of the second part of the ISO 25964:2011 standard that supports establishment of mappings between KOSs with different objectives and structures, has contributed to assessment of the use of mappings in the search and retrieval of information in health area databases. This use of the standard contributed to understanding the results of mapping, with its advantages and limitations, as well as for characterization, evaluation and selection of equivalences to create consistent search strategies in the context of a certain database.

References

Andrade, Julietti. 2015. *Interoperabilidade e mapeamentos entre sistemas de organização do conhecimento na busca e recuperação de informações em saúde: estudo de caso em ortopedia e traumatologia*. Tese doutorado. School of Communications and Arts. University of São Paulo.

Association for Library Collections & Technical Services. 2000. *Task Force on Metadata: Final Report*. Chicago: ALCTS. http://www.libraries.psu.edu/tas/jca/ccda/tf-meta6.html

Bourne, Charles P. 1977. Frequency and Impact of Spelling Errors in Bibliographic Databases. *Information Processing and Management* 13: 1-12.

BSI Group. 2007. *BS 8723-4 Structured Vocabularies for Information Retrieval Guide Interoperability between Vocabularies*. London: British Standards Institutions.

Elsevier. 2013. *Embase Biomedical Answers*. https://www.embase.com/home

International Standard Organization. 2011a. *ISO 25964: Thesauri and Interoperability with Other Vocabularies. Part 1: Thesauri for Information Retrieval*. Geneva: International Standard Organization.

International Standard Organization. 2011b. *ISO 25964: Thesauri and Interoperability with Other Vocabularies. Part 2: Interoperability with Other Vocabularies*. Geneva: International Standard Organization.

Lopes, Ilza Leite. 2002. Estratégia De Busca Na Recuperação Da Informação: Revisão Da Literatura. *Ciência da Informação* 31: 60-71. http://www.scielo.br/pdf/ci/v31n2/12909.pdf

Martinez Tamayo, Ana M., Julia C. Valdez, Edgardo A. Stubbs, Yanina González Terán and M. Inés Kessler. 2011. Interoperabilidad De Sistemas De Organización Del Conocimiento: El Estado Del Arte. *Información, Cultura Y Sociedad* 24: 15-37. http://eprints.rclis.org/17198/

National Information Standards Organization. 2005. *ANSI/NISO Z39.19 – 2005: Guidelines for the Construction, Format, and Management of Monolingual Thesauri*. Bethesda: NISO Press.

U.S. National Library of Medicine. 2009. *UMLS® Reference Manual*. Bethesda: NLM. http://www.ncbi.nlm.nih.gov/books/NBK9676/

U.S. National Library of Medicine. 2013a. *PUBMED*. http://www.ncbi.nlm.nih.gov/pubmed

U.S. National Library of Medicine. 2013b. *Metathesaurus: Mapping Projects: Basic Mapping Project Assumptions*. Bethesda: NLM. http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/mapping_projects/index.html

U.S. National Library of Medicine. 2013c. *UMLS® Metathesaurus® Fact Sheet*. Bethesda: NLM. http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

U.S. National Library of Medicine. 2014. *UMLS Terminology Services*. Bethesda: NLM. https://uts.nlm.nih.gov//metathesaurus.html

U.S. National Library of Medicine. 2015. *Medical Subject Headings (MeSH®)*. Bethesda: NLM. https://www.nlm.nih.gov/pubs/factsheets/mesh.html