IJoL International Journal of Librarianship

ISSN: 2474-3542 Journal homepage: http://journal.calaijol.org

Vocabulary Integration Reexamined

Yunseon Choi

Abstract:

Several thesauri have been published in various domains, or in the same subject domain. This heterogeneity caused the significant incompatibility of transferring or sharing data among different systems and databases. Therefore, thesaurus integration is a solution for handling this incompatibility issue. To achieve interoperability between different thesauri, mapping systems have been developed for establishing equivalents between terms in different thesauri. However, there is still ambiguity in term semantics and hierarchical relations used in thesauri. The purpose of this paper is to reexamine the issues and challenges in vocabulary mapping and integration between different controlled vocabulary systems. The paper outlines the history of the study of vocabulary mapping efforts and suggests a way in which the emerging practices on semantic issues and mapping problems can be articulated.

To cite this article:

Choi, Y. (2018). Vocabulary integration reexamined. *International Journal of Librarianship*, *3*(2), 96-102.

INTERNATIONAL JOURNAL OF LIBRARIANSHIP, 3(2), 96-102

ISSN:2474-3542

Vocabulary Integration Reexamined

Yunseon Choi

Valdosta State University, Valdosta, GA, USA

ABSTRACT

Several thesauri have been published in various domains, or in the same subject domain. This heterogeneity caused the significant incompatibility of transferring or sharing data among different systems and databases. Therefore, thesaurus integration is a solution for handling this incompatibility issue. To achieve interoperability between different thesauri, mapping systems have been developed for establishing equivalents between terms in different thesauri. However, there is still ambiguity in term semantics and hierarchical relations used in thesauri. The purpose of this paper is to reexamine the issues and challenges in vocabulary mapping and integration between different controlled vocabulary systems. The paper outlines the history of the study of vocabulary mapping efforts and suggests a way in which the emerging practices on semantic issues and mapping problems can be articulated.

Keywords: vocabulary mapping, controlled vocabulary, thesaurus, interoperability, ontologies

INTRODUCTION

A thesaurus is a tool for vocabulary control, and it is the most complex type of controlled vocabulary in use in the library and information science professions. Thesauri are often called subject headings in the library context, and generally follow the standards for thesaurus construction using broader term (BT), narrower term (NT), and related term (RT) (NISO, 2017). By guiding indexers and searchers about which terms to use, a thesaurus can help improve the quality of retrieval. Thesauri have been published in various domains, or in the same subject domain. This caused the significant incompatibility of transferring or sharing data among different systems and databases. Thesaurus integration is a solution for handling with this incompatibility issues. Thesaurus reconciliation goes through several processes including mapping, switching, merging, and integration. And thesaurus mapping is a central process for thesaurus reconciliation where terms, concepts and hierarchical relationships between concepts are identified (Doerr, 2001). To achieve interoperability between different thesauri, mapping systems have been developed for establishing equivalents between terms in different thesauri. However, there is still ambiguity in term of semantics and hierarchical relations used in thesauri. This paper reexamines the problems of thesaurus integration and merging, in particular focusing on issues related to vocabulary mapping. This paper provides a review of the challenges and issues of thesaurus

integration and provides a brief account of projects that have investigated the integration and/or merging of controlled vocabularies and different structure in various domains. This paper concludes with a path to exploring the future thinking of research for vocabulary mapping and integration.

CHALLENGES FOR VOCABULARY INTEGRATION

The purpose of thesaurus integration is to integrate various indexes of a collection of documents into a single tool for indexing and retrieving (Aitchison, Gilchrist, & Bawden, 2000). Several thesauri have been published in various domains, or in the same subject domain. UNESCO thesaurus, Getty Art and Architecture Thesaurus (AAT), ERIC Thesaurus, and AGROVOC have been used in digital libraries (Sunny & Angadi, 2017). The examples of thesauri with traditional knowledge include Library of Congress Name Authority file for personal and corporate names and Getty's Thesaurus of Geographic Names (TGN) for place names (Sunny & Angadi, 2017). This heterogeneity caused the significant incompatibility of transferring or sharing data among different systems and databases. Therefore, thesaurus integration is a solution for handling this incompatibility issues.

The main differences in controlled languages in the same field include specificity, exhaustivity, compound terms, synonyms, and inter-relationships (Aitchison, Gilchrist, & Bawden, 2000). For example, there are different levels of specificity and exhaustivity between thesauri in describing the same subjects. In addition, there are different levels of hierarchical structures among thesauri. Doerr (1996) also points out the remaining heterogeneity between different thesauri such as different word use (i.e., language level and terminology degree) and different coverage (i.e., the scope of thesaurus).

Integrating Different Languages and Different Cultures

The central process of thesaurus mapping is establishing vocabulary equivalence (Chan & Zeng, 2002). There are various levels of vocabulary mapping: terminological level (subject heading), semantic level (authority record), and syntactic level (application) (Freyre and Naudi, 2001). However, it is not easy to find one-to-one relationships between terms in different vocabularies due to the differences in linguistic expressions for the same concept. In addition, polysemous terms with multiple meanings hinder vocabularies mapping. On the other hand, the interoperability issues among different thesauri are associated with the questions of integrating the views of different cultures, since controlled vocabulary or subject terms in classification systems need to be properly translated during the process of vocabulary mapping. Literally translated language might be meaningless and there are also difficulties with transferring a whole conceptual structure from one to another culture appropriately (Hudon, 1997).

Integrating Different Structures

Knowledge Organization Systems (KOS) such as subject headings and classification schemes have their own structures and guidelines and differ from one another in their structure, semantic, lexical, and notation or entry features (Iyer and Giguere, 1995). Furthermore, depending on communities and the defined usage of the term, there are a number of ways to investigate vocabularies and their functions (NISO, 2017).

Semantic Problems in Thesaurus Integration

Doerr (2001) points out that hierarchical relations without subsumption (subclass and subproperty) would result in ambiguity in thesaurus mapping and the semantics of hierarchical relations should be made. In thesauri, the semantic differences of hierarchical relations have occurred, because BT

(Broader Term)/NT (Narrower Term) relations were defined differently in different thesauri. In some thesauri it means subsumption (subclass and subproperty), while in other thesauri it can mean BTI (Broader Term Instance) or BTP (Broader Term Partitive). Fisseha (2003) also points out ambiguity in thesauri regrading typical hierarchical relations such as Broader Term and Narrower Term and shows that their semantics are not explicitly defined. For example, "sweet" corn is a property of corn, but in the AGROVOC multilingual thesaurus, the term "sweet corn" is listed as a narrower term of the "Maize" (Figure 1). AGROVOC thesaurus can be accessed online (http://aims.fao.org/standards/agrovoc/functionalities/search). The semantic relationship between terms are not clearly explicated in the thesaurus.

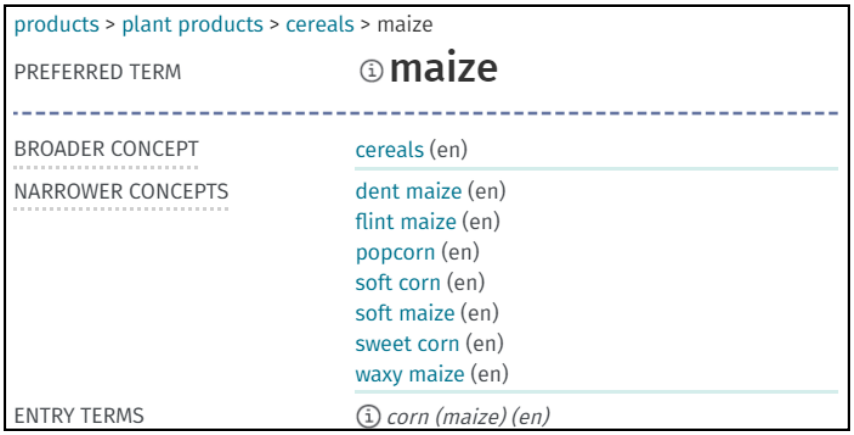

Fig. 1. Hierarchical relations in AGROVOC Multilingual Thesaurus

The subsumption relations between all terms of two thesauri can be identified from a complete mapping using ontological reasoning (Fisseha, 2003). The term *ontology* has been used in several disciplines, from philosophy to computer science. As a branch of philosophy, ontology studies the structures of the objects, properties and relations of reality (Smith, 1997). In computer science, which came out of artificial intelligence, the ontology is a model of the representation of objects in the world with properties and relationships (Garshol, 2004). Gruber (1993) defines an ontology as "a formal, explicit specification of a conceptualisation" (p.1) and explains the definition:

- *Conceptualisation* refers to an abstract, simplified view of the world that we wish to represent for some purpose.
- *Explicit* refers to type of concepts used, and the constraints on their use are explicitly defined.
- *Formal* refers to the fact that an ontology must be able to be read by the computer.
- *Shared* refers to the notion that an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group.

Other researchers describe ontologies as taxonomic hierarchies (Baeza-Yates & Ribeiro-Neto, 1999; Vickery, 1997). Vickery (1997) notes the aspect of taxonomic hierarchies of classes, with class definitions and the subsumption relations. Baeza-Yates and Ribeiro-Neto describe ontologies as hierarchical taxonomies of terms representing topics. Unlike general taxonomies, ontologies classify terms and defines the relationships between the terms, and also expand on taxonomies which clarify the context of terms (NISO, 2017).

In Figure 2, between two terms Maize and Sweet corn, ontological relation such as the "*Kind of*" can be formalized to specify the semantics of relationship between terms. It can be stated that Sweet corn is a *Kind of* Maize.

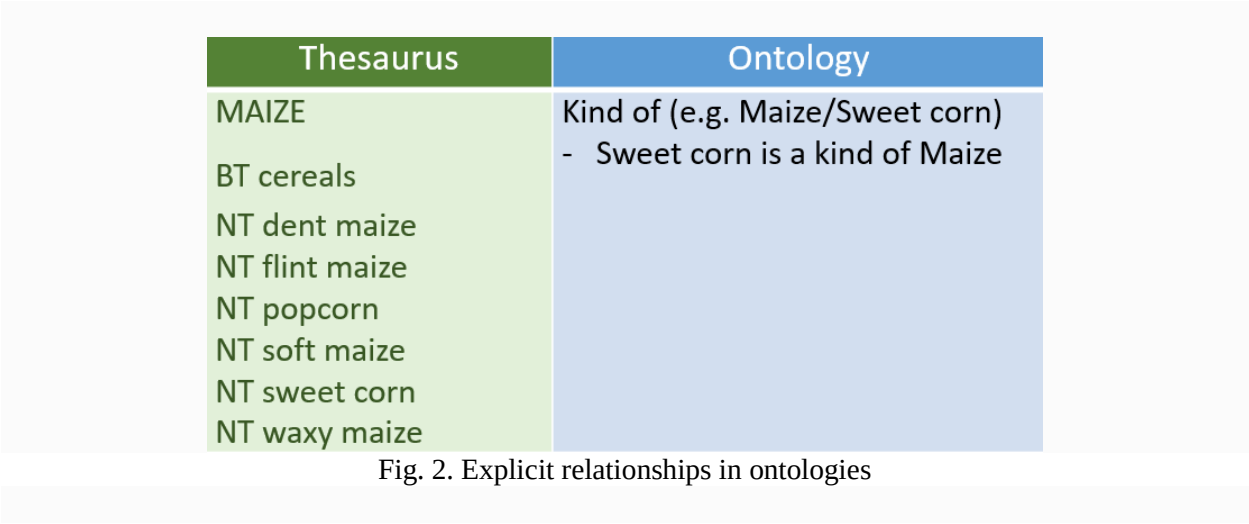

Fig. 2. Explicit relationships in ontologies

Thesaurus integration tools using ontologies and semantic frameworks have been developed to resolve problems associated with ambiguity in hierarchical structure of thesauri. The examples of the tools include the YAM++ Online and the Visual Terminology Alignment Tool (VisTA). The YAM++ Online is a web tool for ontology and thesaurus matching (Bellahsene et al., 2017). The Yam++ Online tool has been partially supported by the French National Research Agency (ANR) within the DOREMUS (Doing REsuable MUSica data) Project focusing on developing controlled vocabularies for music. The Visual Terminology Alignment Tool (VisTA) aims to help users to work on the intellectual handling of the assignment between two terminologies by visualizing vocabulary hierarchies (Axaridou et al., 2018).

EXAMPLES OF MERGED THESAURI

To achieve interoperability among different controlled vocabularies, several attempts to merge thesauri have been made by adopting different approaches and methods to deal with inconsistencies in the process of thesaurus integration:

Integrated energy vocabulary (1979)
The integrated energy vocabulary for the energy domain was developed by integrating 11 existing vocabularies covering the subject area of energy research and development (Niehoff, 1976; Niehoff & Kwansy, 1979).

BRS (Bibliographic Retrieval Services)/TERM vocabulary database (1984)
The BRS/TERM vocabulary database provides natural language synonyms and controlled vocabulary descriptors from seven bibliographic databases in the social and behavioral sciences (Knapp, 1984). TERM database, formerly on the Bibliographic Retrieval Service (BRS), merges terms and codes from controlled languages used in behavioral and social sciences databases.

The National Technical Information Service (NTIS) database (1984)
The National Technical Information Service (NTIS) database is an integrated database from government agencies which have their own thesaurus. The NTIS thesaurus represents a single thesaurus by merging various thesauri and natural language terms (Piternick, 1984).

Unified Medical Language System (1990)
The National Library of Medicine's Unified Medical Language System (UMLS) was developed to integrate biomedical terminology. It aims to build a repository of biomedical terms and their

interrelationships to help users retrieve and organize information (McCray & Hole, 1990). The Unified Medical Language System (UMLS) merged concepts from some 50 sources into a metathesaurus, which retains links to its original sources.

Unified Agricultural Thesaurus (UAT) (1996)

The three producers of the databases such as the US National Agricultural Library (NAL), the UN Food and Agricultural Organization (FAO) and CAB International (CABI) cooperated to create a Unified agricultural thesaurus. It integrated AGROVOC and CAB thesaurus (Clarke & Dextre, 1996) by creating a reorganized, classified structure derived from AGROVOC and CAB thesaurus. It aims to provide users with a comprehensive, multilingual thesaurus system.

HILT (HIgh-Level Thesaurus) (1997)

The HILT ('HIgh-Level Thesaurus') project aimed to focus on the problems associated with cross-searching and browsing by subject in a cross-sectoral and cross-domain environment encompassing libraries, archives, and museums. It integrated distributed and heterogeneous thesaurus databases and merged multilingual and monolingual thesauri (Kramer, Nikolai, & Habeck, 1997).

Precision Medical Vocabulary (2018)

The Precision Medical Vocabulary (PMV) is a controlled vocabulary related with precision medicine (Yu, et al., 2018). It was integrated from several controlled vocabularies including Medical Subject Headings (MeSH), National Cancer Institute Thesaurus (NCIt) and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), and databases in specific domain such as HUGO Gene Nomenclature Committee (HGNC) and Online Mendelian Inheritance in Man (OMIM) for gene, Human Phenotype Ontology (HPO) for human phenotype, DrugBank and RxNorm for drug. The PMV also integrated some biomedical resources such as DrugBank, ClinVar and NCBI Gene with the foundational vocabulary by utilizing a series of mapping and integration strategies.

Controlled Vocabularies for Music Metadata (2018)

Three major French cultural institutions (the French National Library (BnF), Radio France and the Philharmonie de Paris) cooperated to develop controlled vocabularies to describe semantically their catalogs of music works and events. The controlled Vocabularies for Music has been partially supported by the French National Research Agency (ANR) within the DOREMUS (Doing REsuable MUSica data) Project. The Controlled Vocabularies for Music provides music-specific, multilingual controlled vocabularies including topics such as musical genres, keys, or medium of performance (Lisena et al, 2018). The controlled Vocabularies for Music was developed by merging a number of existing vocabularies (IAML, RAMEAU, Diabolo, Itema3, Itema3-MusDoc, and Redomi) and was formalized using Semantic Web languages.

CONCLUSION

As a tool for vocabulary control, thesauri have been used to provide effective access to resources and to achieve indexing consistency. Since several controlled vocabularies have been developed in various domains, the tasks of thesaurus integration and mapping became challenged and faced difficulties due to different culture, different languages, and semantic problems in thesauri. In this paper, we reexamined a set of the issues and trends in vocabulary mapping between different thesauri and aimed to share emerging practices on semantic issues and mapping problems. Term hierarchies in thesauri often do not express subsumption and are ambiguous because they do not

express all subsumption relations. The subsumption relations between terms of two thesauri can be identified from a complete mapping using the successful formalization of ontology and semantic frameworks which result in explicit semantics between terms.

References

Aitchison, J., Gilchrist, A., & Bawden, D. (2000). *Thesaurus Construction and Use: A Practical Manual*. 4th ed. Emerald Group Publishing Limited.

Axaridou A., Konsolaki, K., Kozlov, A., Haase, P., & Doerr, M. (2018). VisTA: Visual Terminology Alignment Tool for Factual Knowledge Aggregation. *Third International Workshop on Semantic Web for Cultural Heritage. In Conjunction with 15th Extended Semantic Web Conference (ESWC 2018),* Heraklion, Crete, Greece, June 3-7, 2018.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*: ACM Press, Addison-Wesley.

Bellahsene, Z., Emonet, V., Ngo, D., & Todorov, K. (2017). YAM++ Online: A Web Platform for Ontology and Thesaurus Matching and Mapping Validation. *The Semantic Web: ESWC 2017 Satellite Events,* Portorož, Slovenia, 2017.

Chan, L. M. and Zeng, M. L. (2002) Ensuring Interoperability among Subject Vocabularies and Knowledge Organization Schemes: a Methodological Analysis. *68th IFLA Council and General Conference*. August 18-24, 2002. Retrieved from http://www.ifla.org/IV/ifla68/papers/008-122e.pdf

Clarke, D. & Dextre, S. G. (1996). Integrating thesauri in the agricultural sciences.' In: Compatability and integration of order systems. *Research Seminar Proceedings of the TIP/ISKO Meeting* (pp. 111-122). Warsaw, 13-15 September 1996. Warsaw: Wydawnictwo SBP, 1996

Doerr. M. (2001). Semantic Problems of Thesaurus Mapping. *Journal of Digital Information, 1*(8)

Doerr, M. (1996) *"Authority Services in Global Information Spaces: A requirements analysis and feasibility study*", Technical Report FORTH-ICS/TR-163, February
English Heritage, National Monuments Record (2000) *NMR Monument Type Thesaurus,* June 19. Retrieved from http://www.rchme.gov.uk/thesaurus/mon_types/default.htm

Fisseha, F. (2003). Reengineering AGROVOC to Ontologies: Step towards better semantic structure. *NKOS Workshop*.

Freyre, E. & Naudi, M. (2001) MACS: Subject access across languages and networks. In Subject Retrieval in a Networked Environment: *Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology,* OCLC, Dublin, Ohio, USA, 14-16 August 2001. Dublin, OH: OCLC.

Garshol, L. M. (2004). Metadata? Thesauri? Taxonomies? Topic Maps! *Journal of Information Science*, *30*(4): 378–391.

Hudon, M. (1997). Multilingual thesaurus construction: integrating the views of different cultures in one gateway to knowledge concepts. *Knowledge Organization 24*(2), 84-91.

Knapp, S. D. (1984). Creating BRS/TERM, a vocabulary database for searchers. *Database, 7*(4), 70–75.

Kramer, R., Nikolai, R., & Habeck, C. (1997) "Thesaurus federations: loosely integrated thesauri for document retrieval in networks based on Internet technologies". *International Journal on Digital Libraries 1*(2), 122-131.

Iyer, H. & Giguere, M. (1995). Towards designing an expert system to map mathematics

classificatory structures. *Knowledge Organization 22*(3/4), 141-147.
Lisena, P., Todorov, K., Ccconi, C., Leresche, F., Canno, I., Puyrenier, F., Voisin, M., Le Meur, T., & Troncy, R. (2018). Controlled vocabularies for music metadata. *19th International Society for Music Information Retrieval Conference,* Paris, France, 2018.
McCray, A. T. & Hole, W. T. (1990). The scope and structure of the first version of the UMLS semantic network. *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care* (pp. 126-130). Los Alamitos, CA: Institute of Electrical and Electronics Engineers, November, 4-7 1990.
National Information Standards Organization. (2017). *Issues in vocabulary management: a technical report of the National Information Standards Organization.* NISO TR-06-2017. Retrieved from https://groups.niso.org/apps/group_public/download.php/18410/NISO_TR-06-2017_Issues_in_Vocabulary_Management.pdf
Niehoff, R. T. (1976). Development of an integrated energy vocabulary and the possibilities for on-line subject switching. *Journal of the American Society for Information Science, 27*(1). 3-17.
Piternick, A. B. (1984). Searching vocabularies: a developing category of online search tools. *Online Review*, *8*(5). 441-449.
Smith, B. (1997). An Essay in Mereotopology. In Hahn, L., ed., *The Philosophy of Roderick Chisholm (Library of Living Philosophers),* LaSalle: Open Court.
Sunny, S. K. & Angadi, M., (2017). Applications of Thesaurus in Digital Libraries. *Journal of Library & Information Technology*. *37*(5). 313-319.
Vickery, B. (1997). Ontologies. *Journal of Information Science*, *23*(4). 277–288.
Yu, M., Liu, Y., Kang, H., Zheng, S., Li, J., & Hou, L. (2018). Building a controlled vocabulary for standardizing precision medicine terms. *KDD2018,* 19-23, August, 2018, London, United Kingdom. Retrieved from http://www.eurecom.fr/en/publication/5646/download/data-publi-5646.pdf

---

About the author

Dr. Yunseon Choi is an assistant professor in the Department of Library and Information Studies at Valdosta State University where she teaches Organization of Information, Thesaurus Construction, Metadata, and Cataloging. She earned her PhD in Library and Information Science from the University of Illinois. Her research interests encompass diverse aspects of information organization in the context of information technologies, including data science with social impacts, linked data, ontologies, and semantic web.