University College London

Department of Information Studies

Discuss similarities and differences of Ontologies and Thesauri. Their role to modern information retrieval systems.

Module leader:

Andreas Vlachidis

This essay is submitted as an assessment for INST0038:

Fundamentals of Information Science

Student ID: 19075478

Word count: 2092

June 4th 2020

TABLE OF CONTENTS

TERMINOLOGY

TERM	DEFINITION
Information Retrieval (IR)	"Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)." *(Manning, et al., 2009).*
Knowledge Organisation System (KOS)	"Knowledge organization system (KOS) is a generic term used for referring to a wide range of items (e.g. subject headings, thesauri, classification schemes and ontologies), which have been conceived with respect to different purposes, in distinct historical moments. They all have in common is that they have been designed to support the organization of knowledge and information in order to make their management and retrieval easier." *(Mazzocchi, 2019).*
Simple Knowledge Organization System (SKOS)	"SKOS—Simple Knowledge Organization System—provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary. As an application of the Resource Description Framework (RDF), SKOS allows concepts to be composed and published on the World Wide Web, linked with data on the Web and integrated into other concept schemes." *(W3C, 2009).*
Knowledge Representation	"Knowledge representation refers to the technical problem of encoding human knowledge and reasoning (Automated Reasoning) into a symbolic language that enables it to be processed by information systems." *(Swain, 2013).*
Controlled vocabulary	"Controlled vocabularies are used to ensure consistent indexing, particularly when indexing multiple documents, periodical articles, web pages or sites, etc. Controlled vocabularies are the broadest category, which includes thesauri and taxonomies. Thesauri and taxonomies are specific kinds of controlled vocabularies, but not all controlled vocabularies are thesauri or taxonomies. Information, that in a simple controlled vocabulary or taxonomy is conveyed through indexing, is embedded into the ontology itself." (American Society for Indexing, n.d.).
Resource Description Framework (RDF)	"Resource Description Framework (RDF) is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed." *(W3C, 2015).*
Ontology Web Language (OWL)	"The Web Ontology Language (OWL) is a language for defining ontologies on the Web. An OWL Ontology describes a domain in terms of classes, properties and individuals and may include rich descriptions of the characteristics of those objects. " *(Bechhofer, 2009).*

1. INTRODUCTION

One of the aspects that people face in the online world is how easily and quickly acquire useful and credible information from areas of their interest. This area of study is called Information Retrieval (IR) which is assisted by Information Retrieval System (IRS) (Yao, 2002). Digital libraries embedded in the environment of the Internet allow their users to access a huge number of information sources - this creates a potentially almost infinite virtual space full of information. The main components of the IRS are the query and indexing system. For the purpose of the IRS, Knowledge Organisation System (KOS) has two aims: the access to knowledge and the use of knowledge (Jain & Singh, 2013). Additionally, KOS works as a bridge between the collection of material and the user's information need (Mazzocchi, 2019) (Hjørland, 2003). A Significant number of KOS including ontology and thesauri have already been implemented in the Semantic Web development through the Simple Knowledge Organization System (SKOS) (Varlan, 2007). Because KOS alone is not able to fully represent knowledge that Semantic Web requires, SKOS serves as a portal for existing KOS to express machine-readable information and allow exchange between software application to the Semantic Web (Peponakis, et al., 2019) (Fernandes, 2015). As mentioned earlier, thesauri and ontology are both one of many elements that belong to KOS. While it is beyond dispute that thesauri is a controlled vocabulary, this should be questioned for ontologies. Multiple studies proved there is not enough comparative knowledge to define the point of "ontology" and "vocabulary" (Kless, et al., 2014) (Almeida, 2013) (Gruninger, et al., 2008). Hence, this discussion will not be part of this essay and vocabularies (thesauri) will be considered as a type of file that embeds information to ontology (American Society for Indexing, n.d.). In it also important to note that even when both thesaurus and ontology are structurally categorised as KOS, thesauri by its function signifies Knowledge Organisation and ontology signifies Knowledge Representation (Fernandes, 2015). Finally, thesauri and ontologies were selected for this essay, which aims to consider overlaps between their functions and effectiveness in the modern IRS.

Initially, the structure of the Semantic Web and its technologies will be discussed. Following to a detailed explanation of ontology and thesauri based on literature. Subsequently, the theoretical comparison of similarities and differences between ontology and thesauri will be explained and conclusion with intention for further research will be examined.

2. SEMANTIC WEB

The term Semantic Web was first introduced in the journal Scientific American in May 2001. The author Tim Berners-Lee pointed out that the information available on World Wide Web (WWW) networks has a completely disorganized meaning and does not guarantee reliability (Berners-lee, et al., 2001). Hence, the Semantic Web was set to become a new evolutionary stage of the existing WWW through guidelines defined by the World Wide Web Consortium (W3C). Semantic Web improves web technologies to exchange and link content by search queries based on the ability to interpret the meaning of words and terms rather than numbers or keywords (Fernandes, 2015). As Figure 1 shows, Semantic Web consists of technological layers, where information is structured and stored according to standardized rules, which makes it easier for machines to find and process (Antoniou & Harmelen, 2004).

The Semantic Web is based on a standardized description of web resources. Each resource is equipped with the same tags and labels, which allows Internet users to work with the WWW as a relational database and query its contents through SQL-like languages (Koivunen & Miller, 2001). This not only helps software agents understand the web page, but also returns the most relevant content back to the user. The emphasis would be on the high accuracy and relevance of the search query response, which already has been proven and validated in the field of information science (Fernandes, 2015). The Semantic Web is empowered primarily by the Resource Description Framework (RDF) and Ontology Web Language (OWL) to represent metadata (Cardiff, 2009) (W3C, 2015). This leads us to the next paragraph where ontology and its modelling languages are discussed.

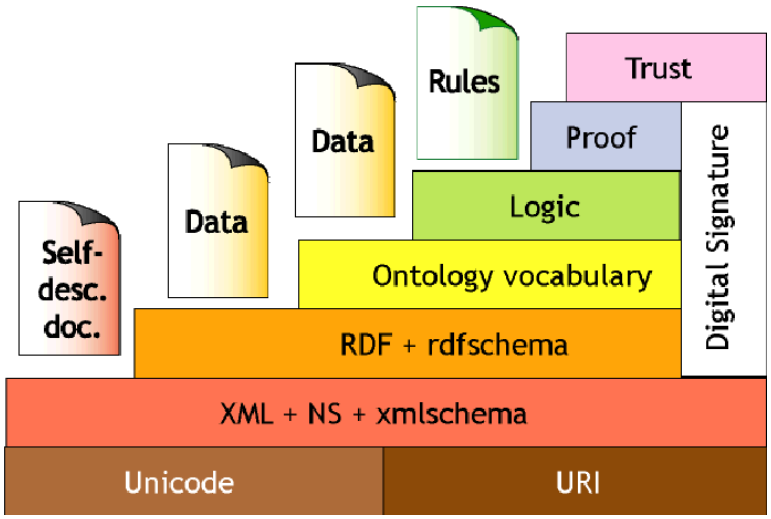

Figure 1: Technological layers of Semantic Web (Patel-Schneider & Siméon, 2002)

3. ONTOLOGY

In computer and information science, an ontology is defined as a populated data model and as a formal and declarative representation that contains a controlled vocabulary (Almeida, 2013) (Gruninger, et al., 2008). Ontology encompasses the definition of entities, properties, relationships and categories between elements of any or all domain of discourse (Gruber, 1995).

Encoding of semantics with the data can be represented in several technologies. For the purpose of this essay, Resource Description Framework (RDF) and Web Ontology Language (OWL) will be examined below.

3.1. RESOURCE DESCRIPTION FRAMEWORK (RDF)

The basic representative language of the Semantic Web is RDF. It is a W3C's vocabulary description language and the most powerful general-purpose knowledge representation framework. RDF is used to describe and model linking structure used in web resources (Amann B., 1999). RDFS extension provides a solid semantic foundation and is used to define and represent the theory of subjects, objects and predicates of the original RDF (S. Decker et al., 2000).

However, the expressive capabilities of RDF and RDFS are very limited. For example, RDF describes resources but provides only a low level of semantics required for metadata statements (Allemang & Hendler, 2011). Due to the great limitations of these languages, it was necessary to create more comprehensive extensions: OWL.

3. 2. ONTOLOGY WEB LANGUAGE (OWL)

The OWL language evolved from DAML + OIL (a combination of DAML and OIL properties). DAML + OIL was such a starting point for the representation of knowledge in the Semantic Web environment while being compatible with XML and RDF. The result was the OWL language for creating web ontologies (Mcguinness, et al., 2002). OWL extension adds more semantics and language richness to RDFS and has become a complementary ontological language that creates solutions to the problems of more complex Semantic Web applications. One of the latest developments by the W3C is OWL2 which is the extension and revision of the original OWL (Allemang & Hendler, 2011) (W3C, 2009). OWL 2 includes easier query capabilities and efficient reasoning algorithms scaled to large datasets. There are several semantic editors that can be used to create ontologies represented in OWL and OWL2, such as Progéte and Jena Softwares have been used to support the development from OWL to OWL2 (Allemang & Hendler, 2011).

In conclusion, OWL, RDF and RDFS have drawn considerable scientific, medical, and commercial attention. As mentioned in the introduction, there is a common opinion that ontology is not considered to be a vocabulary. However, controlled vocabularies such as thesauri are embedded in OWL. Thus, the following paragraph examines the concept of thesauri.

4. THESAURI

The thesauri is one of the most used knowledge organisation tools and one of its traditional purposes is indexing (Kumbhar, 2011). In an IRS context, thesauri constitute to facilitate retrieval, achieve consistency in indexing, and ensure resources are indexed with the same word used by a searcher when formulating a search query (Fernandes, 2015). This description may sound very similar to taxonomy, however, unlike thesauri, it represents mostly hierarchical concepts lacking more complex relationships (Kless, et al., 2014) (Kumbhar, 2011) (American Society for Indexing, n.d.). Thesauri standardization has undergone significant changes in recent times. In 2011 and 2013, two parts of ISO 25964 the international standard for thesauri and interoperability with other vocabularies standards were adopted by the International Organization for Standardization Interoperability (ISO) (Clarke, et al., 2012). With ISO 25964 guidance on managing thesaurus development, thesauri can be created by software such as MultiTES or OSTI Thesaurus software package (Schwarz, 2005).

In addition, ISO 25964 is mainly to support the thesauri creation itself, however, does not include the guidance of thesauri function on the Semantic Web. Therefore, to combine different controlled dictionaries, the language SKOS was created to support publication of thesauri on the web (Peponakis, et al., 2019) (W3C, 2009).

Lastly, the purpose of both ontology and thesauri differs by its nature. To answer the proposed question and logically analyse the results, it is necessary to examine the two specific KOS in the same subject field (domain).

5. THESAURI AND ONTOLOGY: DIFFERENCES AND SIMILARITIES WITHIN THE SAME DOMAIN

This section will analyse specific KOS to identify their similarities and differences. This analysation is based on the fact that ontologies and thesauri are modelled within the same domain (Kless, et al., 2014).

Speaking of the very nature of the specific KOS, thesauri are primarily designed and mainly used to organize concepts and thematic terms, in contrast to ontologies, which are primarily developed to express and organize entities (Peponakis, et al., 2019). If logic-based reasoning from ontologies would be applied to thesauri usage in automatic search expansion, the applied algorithm is much simpler form would result as comparable (American Society for Indexing, n.d.). Hence, the transfer of thesauri into an ontology or vice versa requires complete re-engineering and conceptual re-organisation (Kless, et al., 2014) (Adams, et al., 2012). Specifying a vocabulary in ontology is structurally similar to developing the whole of thesauri (Doerr, 2001).

Because ontologies are modelled in triplets (subject – predicate – object) the role of labels in ontologies match the role that terms have in thesauri (Milton & Daniel Kless, 2010). Resulting in inconsistency checks. However, the meaning of concepts in thesauri overlaps the meaning of concepts in the ontology, because they have different ontological entities (Gruninger, et al., 2008).

The way that ontology and thesauri are aligned horizontally reveals equivalence that hierarchical relationships in thesauri conceptually corresponds to axioms in ontologies (Janowicz & Kessler, 2008) (Fischer, 1998) (Kless, et al., 2014). On the contrary, there is a strong dependency on defining membership conditions and the alignment of top-level ontology in the hierarchical relation to avoiding ambiguity and non-transitivity (Kless, et al., 2014). There are several types of hierarchical relationships, including whole/

part, genus/species, and instance relationships. Generic relationships are the most common as they can be applied in a variety of topics. These might be semantically equivalent to specific KOS but tend to be less consistent than is-a' relationship in ontologies (Tudhope, et al., 2001). However, this also depends on the quality of specific reasoner or defined membership conditions. This might be seen as an improving reference for the generic relationship in thesauri, resulting in better performance in search query expansion.

Terms in thesauri have a different function than labels that are attached to entities in an ontology. Moreover, ontologies do not provide possibilities to distinguish between preferred terms and non-preferred terms (Janowicz & Kessler, 2008). On the other hand, ontologies, unlike thesauri, allow specifying the meaning of the same term through natural language formulation (membership conditions).

Membership conditions express relationships in ontologies and model their concepts, on the contrary thesauri serve mainly navigational and IR purposes and do not contain any scope notes or natural language definitions for their concepts (Kless, et al., 2014). Thus, the expression of a class is more precise in an ontology.

6. CONCLUSION

This essay has reviewed two specific areas: ontology and thesauri, to identify and analyse their differences and similarities in the modern IRS. As the examination demonstrates, both ontology and thesauri are relevant approaches to modelling domains, however, each for different reasons: thesauri are a great tool for Knowledge Organisation and ontologies for expressing rich semantics in Knowledge Representation. Moreover, In the area of Semantic Web, the structure of ontologies was described in detail and with focus on OWL and RDF languages. Also, thesauri were examined with a focus on their usage (indexing and IR) and implementation through SKOS. In addition, this essay was conducted (1) theoretical comparison of ontologies and thesauri, and (2) a comparison specific ontology and a specific thesauri in the same domain. The analysis has shown that the structure and creation of the specific KOS - thesauri and ontologies, should be treated as two different kinds of data sets, but with similar structures. Not only because ontologies are developed for different purposes than thesauri, but also their nature, function and structure is more specific. From a semantic perspective, comparative knowledge between thesauri and ontologies barely exists. It is mainly because ontology is designed to describe objects of reality and thesauri are rather focused on terminology in human language. Moreover, studies have also revealed that ability to fully understand the foundations of what distinguishes thesauri from ontology, needs a detailed understanding of the complex philosophy, structure, nature, semantics and the application of ontologies and thesauri in the field of Semantic Web. As the Semantic Web evolves, it will be fascinating to see what development comes next.

7. BIBLIOGRAPHY

Adams, D., Jansen, L., Lindenthal, J. & Wiebensohn, J., 2012. A method for re-engineering a thesaurus into an ontology. s.l., s.n.

Allemang, D. & Hendler, J., 2011. Semantic Web for the Working Ontologist. 2nd Edition ed. s.l.:s.n.

Almeida, M. B., 2013. Revisiting ontologies: A necessary clarification. Journal of the American Society for Information Science and Technology, 64(8), pp. 1682-1693.

Amann B., F. I., 1999. Integrating Ontologies and Thesauri to Build RDF Schemas. Springer, Berlin, Heidelberg: Research and Advanced Technology for Digital Libraries.

Antoniou, G. & Harmelen, F. V., 2004. A Semantic Web Primer. s.l.:The MIT Press.

Bechhofer, S., 2009. OWL: Web Ontology Language. Boston: Springer.

Berners-lee, T., Hendler, J. & Lassila, O., 2001. The Semantic Web. SCIENTIFIC AMERICAN, 284(5).

Cardiff, J., 2009. The Evolution of the Semantic Web. s.l., Institute of Technology Tallaght, Dublin, Ireland.

Clarke, D., Zeng, S. G. a. & Lei, M., 2012. From ISO 2788 to ISO 25964: The evolution of thesaurus standards towards interoperability and data modelling. Information Standards Quarterly, 24(1).

Doerr, M., 2001. Semantic Problems of Thesaurus Mapping. Journal of Digital Information, 1(8), pp. 1-15.

Fernandes, S., 2015. thedigitalgroup. [Online]
Available at: https://blog.thedigitalgroup.com/
[Accessed 25 5 2020].

Fischer, D., 1998. From Thesauri towards ontologies?. Advances in Knowledge Organization, Volume 6, pp. 18-30.

Gruber, T. R., 1995. Toward principles for the design of ontologies used for knowledge sharing?. International Journal of Human-Computer Studies, 43(5-6), pp. 907-928.

Gruninger, M. et al., 2008. Ontology Summit 2007 – Ontology, taxonomy, folksonomy: Understanding the distinctions. Applied Ontology, 3(3), pp. 191-200.

Hjørland, B., 2003. Fundamentals of Knowledge Organization. pp. 87-111.

Indexing, A. S. f., n.d. Taxonomies-sig.org. [Online]
Available at: https://www.taxonomies-sig.org/about.htm
[Accessed 25 5 2020].

Jain, V. & Singh, M., 2013. Ontology Based Information Retrieval in Semantic Web: A Survey. International Journal of Information Technology and Computer Science, 5(10), pp. 62-69.

Janowicz, K. & Kessler, C., 2008. The role of ontology in improving gazetteer interaction. International Journal of Geographical Information Science, 22(10), pp. 1129-1157.

Kless, D., Milton, S., Kazmierczak, E. & Lindenthal, J., 2014. Thesaurus and ontology structure: Formal and pragmatic differences and similarities. Journal of the Association for Information Science and Technology, 66(7), pp. 1348-1366.

Koivunen, M.-R. & Miller, E., 2001. Semantic Web Kick-off Seminar in Finland. Finland, s.n.

Kumbhar, R., 2011. Library Classification Trends in the 21st Century. 1st ed. Oxford: Chandos Publishing.

Manning, C. D., Raghavan, P. & Hinrich Schütze, 2009. Introduction to Information Retrieval. s.l.:Cambridge University Press.

Mazzocchi, F., 2019. ISKO. [Online]
Available at: https://www.isko.org/cyclo/kos
[Accessed 26 5 2020].

Mcguinness, D. L., R. Fikes, J., Hendler & L. A. Stein, 2002. DAML+OIL: an ontology language for the Semantic Web. EEE Intelligent Systems, 17(5), pp. 72-80.

Milton, S. & Daniel Kless, 2010. Comparison of thesauri and ontologies from a semiotic perspective. s.l., Australian Computer Society.

Patel-Schneider, P. F. & Siméon, J., 2002. Building the Semantic Web on XML. The Semantic Web — ISWC 2002, pp. 147-161.

Peponakis, M., Mastora, A., Kapidakis, S. & M., D., 2019. Expressiveness and machine processability of Knowledge Organization Systems (KOS): an analysis of concepts and relations.. International Journal on Digital Libraries, 20(4), pp. 433-452.

S. Decker et al., 2000. The Semantic Web: the roles of XML and RDF. EEE Internet Computing, 4(5), pp. 63-73.

Schwarz, K., 2005. Domain model enhanced search - A comparison of taxonomy, thesaurus and ontology. s.l.:Master of Content and Knowledge Engineering University of Utrecht.

Swain, M., 2013. Knowledge Representation. Encyclopedia of Systems Biology: Springer, New York, NY.

Tudhope, D., Alani, H. & Jones, C., 2001. Augmenting thesaurus relationships: possibilities for retrieval. ournal of Digital Information, 1(8).

Varlan, S. &. T. C., 2007. SIMPLE KNOWLEDGE ORGANISATION SYSTEM. Supplement Proceedings of CNMI 2007, pp. 299-308.

W3C, 2009. W3. [Online]
Available at: https://www.w3.org/
[Accessed 26 5 2020].

W3C, 2015. [Online]
Available at: https://www.w3.org/standards/semanticweb/ontology
[Accessed 27 5 2020].

Yao, Y. Y., 2002. Information retrieval support systems. 2002 IEEE World Congress on Computational Intelligence. .