

Lists, Taxonomies, Lattices, Thesauri and Ontologies: Paving a pathway through a terminological jungle

V. Pieterse and D.G. Kourie

Department of Computer Science, University of Pretoria, Pretoria, South Africa

1. Introduction

Knowledge organization systems (KOSs) are mechanisms for organizing information - they are at the heart of every library, museum and archive. The organization of knowledge also plays a key role in all scientific research aimed at knowledge creation. The acquisition of new knowledge depends on the ability to create new categories through the discovery of new patterns and new relationships.

Owing to the need to organize information in nearly every discipline, KOSs with varying attributes, content and structures have been developed independently in different domains. These scattered and independent developments have given rise to a conglomeration of classification-related terms which tend to be used inconsistently in some research fields, and even more so across different research fields. Jacob (2010) argues that there is a need to assess the terminology of classification, and for a methodical and critical assessment of the criteria governing the development and implementation of classifications.

Misunderstanding is caused by the way different domains, or different parts of the same domain, overload terms by ascribing different meanings to the same terms (McComb 2003). This seems to be particularly prevalent in information and knowledge management. According to Noy and

McGuinness (2001), authors often use definitions that contradict previously cited definitions. Gilchrist (2003) states that the meanings of the words thesaurus, ontology and taxonomy seem to overlap significantly and are moreover used contradictorily. Frank (2006) points to an irony: “It is amazing that the field of ontological studies that pretends to clarify the meaning of words is itself entangled in a confusing terminology”. Hilera et al. (2011) assert that this situation may change, owing to the publication of works in which the development of ontologies is related to the development of classic terminological tools such as vocabularies, taxonomies and thesauri. Nevertheless, confusion about the meaning of these terms prevails.

Research has been conducted over a wide spectrum to create a general semantic web that comprises all knowledge. On the one end of the spectrum are researchers in Library and Information Science (LIS) who are interested in knowledge organization with the aim of more effective knowledge retrieval. On the other end of the spectrum computer scientists conduct research to advance Artificial Intelligence (AI). They are interested in building intelligent systems with a main focus on knowledge discovery. For an ideal Semantic Web to be created, these approaches have to embrace each other. AI researchers have to incorporate the vast body of semantic information that is embedded in existing classifications, whereas LIS researchers have to allow structural revisions and amplifications of their systems to facilitate higher levels of interoperability between diverse systems.

In seeking to bridge the terminological gap between the researchers and practitioners of opposing disciplines, we identified five broad classes of KOSs, which we call lists, taxonomies, lattices, thesauri and ontologies. We provide our definitions for these terms and explain their meaning

and usage. We discuss a few commonly used terms, namely catalogue, index, lexicon, knowledge base and topic map. We delineate how each of these relates to one another and identify which type of KOS best suits the implementation of each of these.

2. Classification of knowledge organization systems

Diverse kinds of collections, employing a variety of data models, have evolved over the years. The terminological preferences of practitioners in each domain have resulted in the use of terms that have different meanings in different contexts. Further confusion arises when KOSs are classified according to different criteria; classification of KOSs based on purpose will differ from classification of KOSs based on content, and from classification of KOSs based on structure. In agreement with Jacob (2004) who emphasizes that the organizational structure of a collection dictates how it can be used, we offer a classification of KOSs based on the inherent structure and types of content of KOSs. Our classification is intended to augment other classifications in an attempt to resolve ambiguities and to create a shared vocabulary in support of better communication and collaboration.

Kempf et al. (2014) point out that the development of matching systems is impeded by the different formats that are used to represent KOSs. For this reason a classification that focuses on the structure of the underlying data instead of the content or purpose, may be more suited when the aim is to integrate the disparate classificatory systems that exist in our increasingly connected environment. Such integration is a necessary step towards achieving the Semantic Web aspiration (Green, 2006).

2.1 Hodge's classification

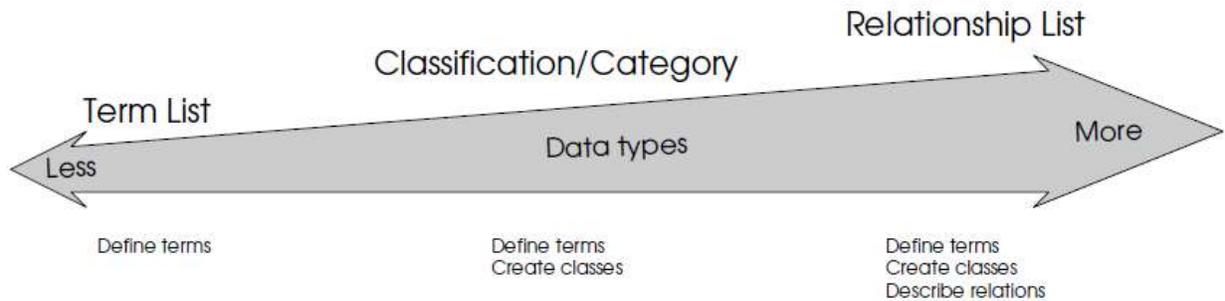


Figure 1: Increasing data types included in KOSs

In her report addressing knowledge organization systems, Hodge (2000) describes a classification of KOSs based on characteristics such as structure and complexity, relationships between, and historical function. This classification is presented in Figure 1. Three broad categories are identified: term lists, classifications and categories, and relationship lists. A term list is the simplest form of a KOS. It lists and defines terms. Examples of term lists are authority files, glossaries and dictionaries. If a KOS has a hierarchical structure it is more than a list and is therefore identified as a classification or category. Subject headings such as the Medical Subject Headings (MeSH) and the Library of Congress Subject Headings (LCSH) belong here along with more comprehensive schemes such as the Dewey Decimal Classification (DDC) (Dewey 2011), the Universal Decimal Classification (UDC) and the latest Bliss Bibliographic Classification (BC2), as well as a wide spectrum of formal and less formal taxonomies. These all provide hierarchical arrangements of topics. At the high end of Hodge's classification of KOSs is relationship lists. These are the more comprehensive systems which emphasize the connections between terms and concepts. These connections are relations such as those defined in the ANSI/NISO Z39.19 (2005) standard for thesaurus construction, as well as relations found in

semantic networks and ontologies. The types of relations found in semantic networks and ontologies may go beyond the types of relations typically found in thesauri.

2.2 Gilchrist's classification

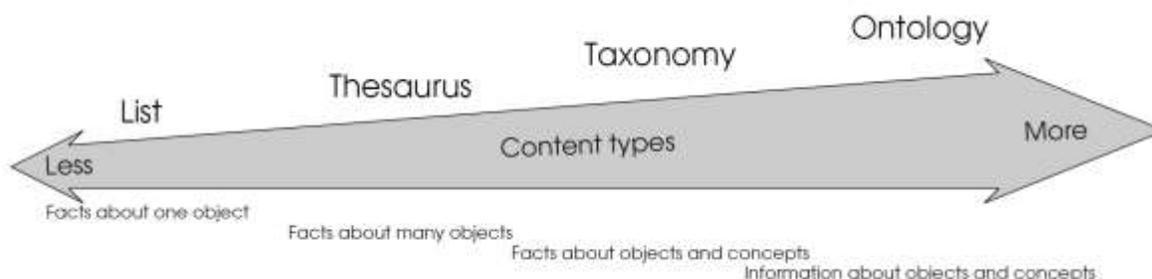


Figure 2: Increasing types of content included in KOSs

Gilchrist (2001) clarifies the distinction between different KOSs. He proposes the progression on content types shown in Figure 2. This progression is based on the history and domain of usage of the different KOSs. Lists - such as the back-of-book indices - appear at the low end. Next are thesauri that are applied to keep facts about things such as books, documents and artifacts with the intention to support accessibility to these items in libraries and museums. Thesauri mostly index physical items. Taxonomies have expanded thesauri because it has become increasingly necessary to store facts about topics rather than merely storing facts about physical things; consequently, taxonomies include broader concepts than thesauri. The next progression is the enhancement of the type of detail that is kept per concept. Ontologies allow the modeling of information about concepts as opposed to being a collection of facts about these concepts. Information such as definitions and notes on usage, syntactic strings and phrases, as well as morphological variations increases the ability to support the application of intelligent processes to improve the accuracy of retrieval.

2.3 Our classification

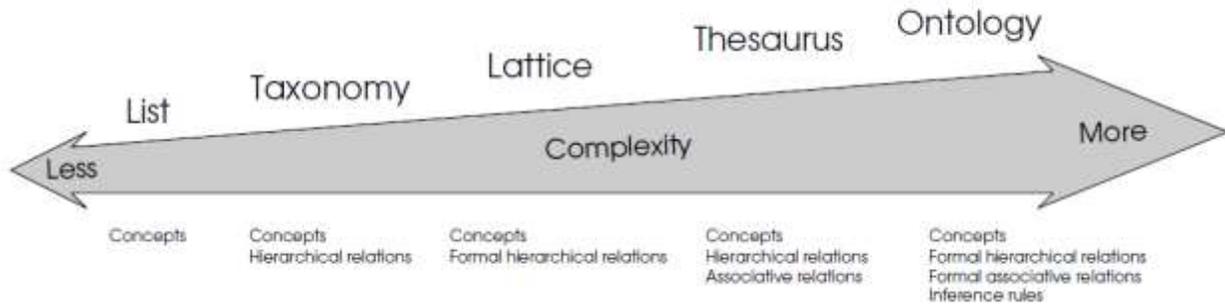


Figure 3: Increasing structural complexity of KOSs

In our classification of KOSs we consider the inherent structure of classifications. Classes of KOSs are characterized by the progressive addition of features that enhance the capabilities offered by these KOSs. The addition of these features contributes to their increased complexity. We call these classes of KOSs lists, taxonomies, lattices, thesauri and ontologies. This progression is summarized in Figure 3. Lists are found at the simplest end. The addition of hierarchical relationships in taxonomies enables more advanced retrieval processes which can make use of broader and narrower terms to improve recall and precision respectively. The next class of KOSs is lattices. These are hierarchical structures encoded as formal concept lattices. This formalization allows for computations that have the potential to improve the precision and recall when information is retrieved using these computations. A further enhancement offered by thesauri is the inclusion of semantic relationships beyond hierarchical relationships. These relationships are intended to contribute to the reasoning power that is to be built into applications that use thesauri. The final enhancement extends KOSs beyond controlled vocabularies to

ontologies. This enhancement entails two things: firstly, the addition of inference rules in the form of meta-relations, constraints, conditional rules or production rules, and secondly, the formalization of its content.

2.4 Comparison

These classifications agree at the lower end. Our concept of a list correlates with “lists” as defined by Gilchrist and with “term lists” in Hodge’s classification. Our concept of a taxonomy correlates with “Classifications and Categories” as defined in Hodge’s classification, but ours is narrower than the “Taxonomies” in Gilchrist’s classification. The only difference between a taxonomy and a lattice in our classification, is the higher level of mathematical rigour applied in lattices. Neither Hodge nor Gilchrist makes any distinction between KOSs based on the methods applied when manipulating the data in the KOS. For this reason lattices will be classified in their classifications in the same category as taxonomies would. Our concept of a thesaurus correlates with the “relationship lists” defined by Hodge as KOSs containing information about relations between topics. Hodge classifies thesauri as well as semantic networks and ontologies as “Relationship Lists”. Thus, Hodge’s concept of semantic networks and ontologies correlates with our concept of a thesaurus. Our concept of a thesaurus also includes thesauri, taxonomies and ontologies as defined in Gilchrist’s classification. Our concept of an ontology correlates with its formal meaning as used in computer science. It refers to KOSs that extend both Hodge’s and Gilchrist’s concept of an ontology. Consequently, the type of KOS we call an ontology is not included in their classifications.

Other classifications of KOSs are offered by Hedden (2010), Lambe (2007), Borgo and Leitão (2007) and Pepper (2010). These are different from one another and also differ from the previously discussed classifications. Hedden (2010) and Lambe (2007) call all KOSs taxonomies and differentiate between different kinds of taxonomies, Borgo and Leitão (2007) call them all different types of ontologies, while Pepper (2010) explains the same progression but classifies them as indices, thesauri and topic maps.

In the following sections we describe the five broad classes of KOSs in our classification.

3. List

The simplest structure of a KOS is a linear list of related things, together with some descriptions of and/or properties owned by these things.

3.1 Definition

A list is a linearly organized collection that contains items and their attributes.

3.2 Discussion

A list as defined here is commonly called a “flat file” in computer applications. An item on a list correlates with a record in a file and its attributes are the fields of the record. A list does not include a deep organization or complex structure. The attributes are some properties and/or descriptions of the items. An item may have any number of attributes. The attributes may vary from simple values and identifying words to in-depth descriptions.

The order in which the items are presented on a list has no particular significance. Most often such lists are sorted in a logical order that simplifies retrieval; for example, items may be presented in numerical or in alphabetical order according to one of its attributes.

Lists in printed as well as in electronic form are given different names depending on the content and structure of the items on the list, as well as the objective of the list. Terms such as catalogue, authority file, dictionary, glossary, gazette and encyclopedia may be used to refer to different types of lists. The distinction between different types of lists arises from differences in the content and purpose of the various lists; however, their structure complies with our definition of a list regardless of their content.

In practice, very few KOSs are pure lists. As soon as a list structure is augmented by a higher form of organization, such as the grouping together of related items or the addition an index to the items on the list, the augmented structure is no longer purely a list.

4. Taxonomy

The word taxonomy comes from the Greek τάξις, (taxis, meaning order) and νόμος, (nomos, meaning law or science) (Wikipedia 2008b). The Webster dictionary has included the word taxonomy since its 1828 version (OED 1989). It originated, however, much earlier as a term that is used to refer to the systematic categorization and naming of living organisms. Carl Linnaeus, known as the father of modern taxonomy, published a first edition of his *Systema Naturae* in the Netherlands in 1735 (Wikipedia 2008a) and then already used the term.

4.1 Definition

A taxonomy is a hierarchically organized collection that contains items and their attributes.

4.2 Discussion

A taxonomy is created by grouping things in a domain into categories and sub-categories. Often sub-categories are formed several levels deep. When two concepts are in a hierarchical relationship, the super-concept is called the hypernym of the sub-concept, and the sub-concept is called the hyponym of the super-concept. It ensures the transfer of properties from super-concepts to sub-concepts. Sometimes restrictions are placed on the hypernyms and hyponyms in the taxonomy. Garshol (2004) describes faceted classification as a disciplined way to construct a thesaurus. We conversely call such a faceted hierarchical structure a taxonomy because it only has hierarchical relations.

When each concept is limited to only one hypernym, the taxonomy is strictly hierarchical and can be presented in a tree structure. In such a taxonomy the position of every item is uniquely determined as each item may only belong to one sub-class which in turn may only have one super-class. If concepts are permitted to have multiple hypernyms, the resulting structure is a semi-lattice as defined in mathematical lattice theory (Birkhoff 1948; Wille 1982; Grätzer 2011). Lambe (2007) calls them poly-hierarchies.

Green (2011) compared the hierarchical relations that are encoded in the 23rd edition of the DDC system with the different kinds of hierarchical relations that can be found in existing classification systems that adhere to the ANSI/NISO Z39.19 (2005) guidelines, as well as the hierarchical relations that can be expressed using the Web Ontology Language (OWL2, 2009).

She observed that OWL is surprisingly weak when compared with the ANSI/NISO standard in terms of distinction between different types of hierarchical relations. She also observed that the DDC is weaker than the ANSI/NISO standard to encode the rich variety of types of hierarchical relations that are supported by the DDC.

5. Lattices

In mathematics a lattice is a non-empty, partially ordered set along with two binary operations that are idempotent, commutative and associative, and satisfy the absorption law. The study of lattices is called lattice theory. Lattices offer a natural way to formalize and study the hierarchical ordering of objects. When a taxonomy is formally structured by describing the relations between objects in terms of attribute sets and manipulated by applying lattice operations, it is classified as a lattice in our classification of KOSs.

5.1 Definition

A lattice is a hierarchically organized collection that contains items and their attributes in which these items and their attributes are formally presented as a concept lattice.

5.2 Discussion

The only difference between a taxonomy and a lattice, as defined here, is the higher level of mathematical rigour applied in lattices. When the information in a taxonomy is represented in a lattice it can be manipulated using formal concept analysis (FCA). FCA is a mathematical formalization of the concepts “concept”, “concept extension”, “concept intention” and “conceptual hierarchy” according to the classical theory of concepts attributed to Aristotle. FCA

finds practical application in many fields, including artificial intelligence, knowledge management and software development. According to Priss (2006), it is as fundamental to hierarchies and object/attribute structures as relational algebra is to relational databases.

A specific object in a lattice may belong to multiple classes and each sub-class may have a number of super-classes. If each pair of objects in a structure has a unique connection called its “join” (i.e. least upper-bound), the structure is called a semi-lattice. A complete lattice is a semi-lattice with the additional stipulation that each pair of items also has a unique connection called its “meet” (i.e. greatest lower-bound). A complete lattice is equipped with an algebraic structure that allows for computation (Stumme 2002). Valtchev et al. (2003) developed an open platform to support lattice manipulation. The application of lattice manipulation in an information system that stores its data in a complete lattice allows for the creation of fairly sophisticated systems through the application of formal concept analysis (Carpineto & Romano 2004; Cimiano et al. 2004; Laukaitis & Vasilecas 2007).

The principle of duality that exists between the extent and intent of the formal concept forms a Galois connection between the two. The existence of a Galois connection between the power set of all concepts and the power set of all attributes implies that if one adds an object to an object set, the intent of the resulting object set can only be a subset of the intent of the original object set, or at most, be the same set as the intent of the original object set. Kumar et al. (2013) contend that this principle makes FCA particularly suitable for IR applications, since a smaller set of keywords returns a larger document set and a larger set of keywords returns a smaller document set. Bo Ning (2013) showed that a text information retrieval algorithm that uses

vectors and concept lattice theory performs better with regard to accuracy and response than methods based on full-text retrieval and pure similarity retrieval.

6. Thesaurus

The word thesaurus is a Latin word that is the latinization of the Greek word Θησαυρός (thesaurus, meaning treasure store) (Wikipedia 2011). The term, however, is currently more often used to refer to a classified list of terms and their synonyms in a particular field. This change in meaning from a treasure collection to a dictionary of synonyms was brought about by the publication of Roget's *Thesaurus of English Words and Phrases* (1912). Most likely, the meaning of the word thesaurus in this title was chosen to describe the collection as a valuable resource of words and phrases for practical application. When this book was first published in 1852, Roget described it as a "classed catalogue of words". This book was widely used and its title gradually became synonymous with its intent, just as the word google is currently used as a verb synonymous with the concept of searching the internet, whereas it is the domain name of a popular website offering web-searching functionality.

Wiktionary (2010b), a wiki-based open-content dictionary, defines a thesaurus as a "publication, usually in the form of a book that provides synonyms (and sometimes antonyms) for the words of a given language."

Spärck-Jones (1992) pointed out that the word thesaurus is used in three senses. Most commonly it is used to refer to a vocabulary reference work to aid writing. This usage complies with the above definition offered in Wikipedia. We use the term *Vocabulary Reference Thesaurus* (VR

thesaurus) to specify this kind of thesaurus. Secondly, in the Library and Information Science (LIS) field, a thesaurus is a technical term that refers to an instrument that is used for vocabulary control in order to achieve consistency in the description of items and to facilitate retrieval. A thesaurus in this sense does not define or explain the concepts of the domain, but rather focuses on describing the relations between the concepts and uses scope notes mainly to delineate the meaning of concepts and to resolve ambiguities. We use the term Information Retrieval Thesaurus (IR thesaurus) to specify this. Lastly, in the Artificial Intelligence (AI) field the term refers to the data store of words and phrases that is commonly used as a natural language processing resource, hence the term Natural Language Processing Thesaurus (NLP thesaurus). In the following definition we capture the commonality of the structure of a thesaurus that is relevant for all the above-mentioned uses of the term.

6.1 Definition

A thesaurus is a collection that contains items within a selected domain. A thesaurus allows for the specification of the attributes of items as well as the definition of equivalence, hierarchical, associative and/or contrast semantic relations between its items.

6.2 Discussion

In our classification a thesaurus can be distinguished from a taxonomy or lattice by referring to the types of relations that are definable in the KOSs. Taxonomies and lattices accommodate only hierarchical relations (hypernymy and hyponymy), whereas thesauri additionally allow a number of other types of relations.

It is important to note that not all thesauri support the same types or granularity of different types of relations. A thesaurus complying with the ANSI/NISO Z39.19 (2005) standard for thesaurus construction for information retrieval fits this definition. Our definition, however, permits a wider range of relations as well as more detailed content than permitted by this standard. We endorse Gilchrist's (2003) statement that "there is no intrinsic reason why a conventional thesaurus should not be extended and elaborated to include, for example, term definitions, notes on term usage, and more explicitly defined relations."

KOSs like terminological databases and lexical databases are sometimes seen as different from thesauri because of their use of finer-grained relations between items. As we do not refer to the granularity of the relations in our definition of a thesaurus, we classify all these KOSs as thesauri. Our definition of a thesaurus also correlates with the definition of other terms used in the computer science domain for similar data models. Examples are conceptual schemata for databases and domain models in software engineering. Of course, these entities are not referred to as thesauri in their respective contexts. An understanding of the synonymy of the concept thesaurus as we defined it, and the concepts of the models and schemata used in other disciplines can possibly promote synergies.

The description of the relations between the items contained in a thesaurus is its most vital characteristic. These relations represent semantic relationships that may exist between concepts. As indicated in our definition, the types of semantic relations that can be specified between items in a thesaurus can be classified into four main types: equivalence, hierarchical, associative and contrast. It is important to note that not all types of thesauri include all types of relations. The

first three of these are usually mentioned with regard to IR thesauri (Tudhope & Binding 2008), while VR thesauri mainly use equivalence and contrast relations. NLP thesauri usually apply all of the above-mentioned types of relations and also require distinction between different types of relations having a finer granularity than the other types of thesauri.

7. Ontology

The word ontology is derived from the two Greek words *όντος* (ontos, meaning to be) and *λογία* (logia, meaning science, study or theory) (Wikipedia 2010). Ontology is the philosophical study of the nature of being, existence or reality. Ontology focuses on how knowledge can be represented. It aims to determine what entities exist and philosophizes about how these entities can be classified and/or relate to one another. The following comprehensive formal definition of an ontology originating from an AI perspective is supplied by Nguyen et al. (2010):

According to a formalism in conceptual structure theory, an ontology consists of three partially ordered sets of concept types, relation types, and meta-relation types, a set of individuals (which are instances of concept, relation and meta-relation types), and logical rules concerning semantic relationships between those structures, objects and their attributes.

The above description is the basis for our definition of an ontology which incorporates aspects of the concept ontology, as it was formally developed with mathematical rigour, as well as the concept ontology as it is used in LIS.

7.1 Definition

An ontology is an electronically stored collection that comprises a thesaurus combined with a set of inference rules.

7.2 Discussion

Our concept of an ontology correlates with its formal meaning as used in computer science. It refers to KOSs that extend both Hodge's and Gilchrist's concept of an ontology. We define an ontology as an extension of a thesaurus. It contains items representing concepts, their attributes and relations in a more formal structure than required for thesauri in general. There are two aspects that distinguish an ontology from other KOSs, namely the requirement that the representation should be based on a formalism, and that it should include inference rules. These rules are required if the knowledge encoded in the KOS is to be manipulated and interpreted by a computer program. In AI circles a program that manipulates the data in a KOS through the application of inference rules is called an inference engine. The semantic expressiveness of ontologies exceeds that of other KOSs because of the availability of inference rules. They also tend to contain more detailed information about concepts, deeper hierarchical levels of concepts, and richer relationships between concepts.

Hedden (2010) calls a classification system with relationships that bear meaning an ontology. Noy and McGuinness (2001) equate an ontology with a conceptual schema. Furthermore, Grefenstette (in Buitelaar & Cimiano 2008, xi) states that an ontology of the semantic web is a data structure containing a domain vocabulary and the relations between the elements in that domain. Like Hodge and Gilchrist, these authors also use the term ontology to refer to KOSs that

we insist on classifying as thesauri. We concede that terming these kinds of collections ontologies resolves the ambiguity between the different kinds of thesauri we have mentioned. We have observed that authors tend to use the term thesaurus to refer to a VR thesaurus and sometimes to an IR thesaurus, and to use the term ontology to refer to more comprehensive IR thesauri and to NLP thesauri. While agreeing that these views are acceptable, we prefer to interpret the word ontology as described by our definition. We also refer to the extension of a taxonomy that allows for the definition of non-hierarchical relations between terms as a thesaurus, regardless of the granularity of such relations. When practitioners communicate across disciplines, we caution against the use of the word ontology to refer to collections that are not formalized and do not incorporate inference rules.

The set of inference rules that are an essential part of an ontology, is intended to empower a computer program that is used to manipulate the data in the ontology to such an extent that it can simulate reasoning. A program may use inference rules to deduce new information or to verify the correctness of assertions in order to answer user queries. Inference rules can take many forms depending on the language that is used to express them, as well as the design of the program that applies them. We distinguish four forms of inference rules, namely 1) meta-relations, 2) restrictions such as quantification restrictions on attributes and cardinality restrictions on relations, 3) conditions, and 4) rules to form compound or new concepts.

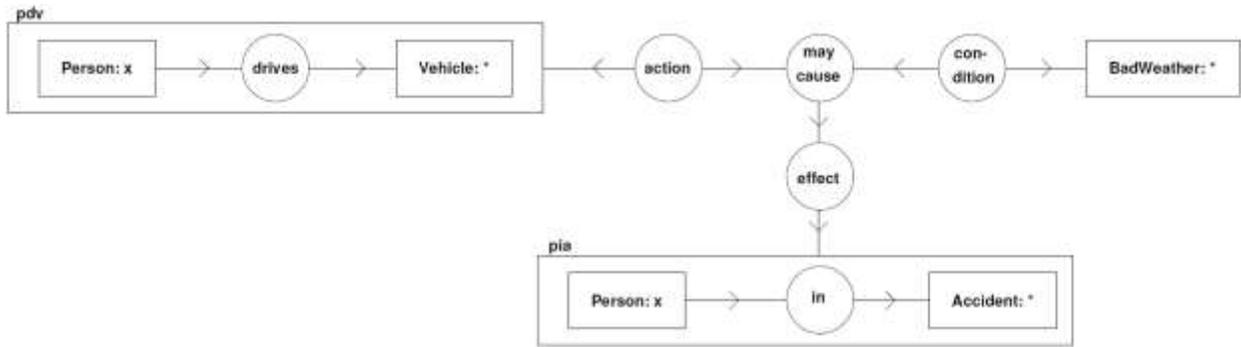


Figure 4: Conceptual graph of an assertion using display form (DF)

Meta-relations are relations between relations. Relations are in essence predicates regarding concepts. If these predicates themselves act as arguments of predicates, a relation between predicates is defined. Such relations are called meta-relations. Figure 4 is a formal representation of an assertion as a conceptual graph (Sowa 1984) using the graphic display form (DF) for conceptual graphs. The assertion represented here is: driving in bad weather may cause an accident. The relation *may cause* is an example of a meta-relation as two of its arguments are relations. This example was given by Nguyen et al. (2009). The acronyms pia and pdv in this figure are labels that may be used to refer to these objects if needed. These labels are defined and applied in a different representation of this assertion in Figure 7.

```
[[Person] - (drives) --> [Vehicle]], [Bad weather] -
(may cause) --> [[Person] - (involved in) --> [Accident]].
```

Figure 5: Conceptual graph of the assertion in Figure 4 using linear form (LF)

Conceptual graphs can be represented in several different concrete notations. Figure 5 is a representation of the same assertion as a conceptual graph using the readable linear form (LF) for conceptual graphs.

Quantification restrictions on attributes of concepts as well as cardinality restrictions in the relationships between concepts are also considered to be inference rules (Kang et al. 2009). They can be used to verify the correctness of assertions or eliminate incorrect results when deductions produce multiple consequences. Most of the concept languages proposed for the implementation of ontologies allow for the definition of such restrictions.

Conditional statements are the most common kind of inference rule. They often appear as expressions in the form of if-then statements. They are used to express sets of actions and heuristics (Corcho & Gómez-Pérez 2000). Such inference rules often appear as axioms that are expressed through logical expressions (Gruber 1993). Different formalisms such as propositional calculus, description logic (DL), first order predicate logic (FOPL) and second order predicate logic (SOPL) may be used.

$$P \wedge (P \Rightarrow Q) \Rightarrow Q \quad (1)$$

$$\frac{P, P \vdash Q}{Q} \quad (2)$$

Figure 6: Different notations expressing the modes ponens rule

Equations 1 and 2 in Figure 6 express the modes ponens rule using different notations. The argument form of this expression has two premisses. The one premiss is the if-then or

conditional claim, namely that P implies Q. The other premiss is that P, the antecedent of the conditional claim, is true. From these two premisses, it can be logically concluded that Q, the consequent of the conditional claim, must be true as well. Equation 1 expresses this rule as a propositional calculus statement, while Equation 2 is an expression using FOPL rule notation. In AI, modus ponens is often called forward chaining. Many other first order logic axioms are explained in Sakharov (n.d.).

Production rules constitute another class of inference rules that produce new concepts. Rules by which compound concepts may be formed are the simplest form of production rules. An ontology that contains such rules is generative and only needs to maintain a set of atomic concepts. When it has to deal with compound concepts, the rules to form compound concepts are applied.

Practitioners transform existing thesauri to ontologies in the hope that they can be used to facilitate automatic reasoning. This transformation on a structural level requires two things: firstly formalizing the data using some standard and secondly, the addition of inference rules. This transformation is a complex process. Green et al. (2013), for example, experienced several difficulties in the interpretation of existing knowledge encoded in the DDC system. Furthermore, several technologies have been proposed for the encoding of ontological information, such as the Resource Description Framework (RDF) (Klyne & Carroll 2004), Web Ontology Language (OWL) (McGuinness & van Harmelen 2004), and the Simple Knowledge Organization System (SKOS) (Miles & Bechhofer 2009), to name but a few. The ultimate aim is the interoperability of all systems. This can only be achieved if all apply the same standard technology. It is unfortunate

that a powerful and established standard such as the ISO/IEC 13250 (2003) is not more widely adopted.

7. Familiar knowledge organization systems

In this section we discuss some widely used types of KOSs. We show how they relate to one another and describe their data models in terms of the five classes of KOSs we have defined.

7.1 Catalogue

According to the Merriam-Webster online dictionary (Merriam-Webster 2010a), a catalogue is a complete enumeration of items arranged systematically with descriptive details. An example is a printed telephone directory that contains names, addresses and phone numbers. In its simplest form a catalogue is classified as a list.

The term catalogue is sometimes used to refer to a larger collection that may have a kind of hierarchical structure. For example, Chaffey et al. (2009) define a catalogue in the context of the organization of information on the internet as an interface that provides a structured listing of registered web sites in different categories. When a catalogue has sub-divisions it can no longer be classified as a list; instead, it has to be a simplistic, usually hierarchic, taxonomy.

7.2 Dictionary

The word dictionary is of Medieval Latin origin. It was created by combining two words: dicto (meaning speaking) and arium (meaning room). It literally means a place of words. According to Wiktionary (2010a), a dictionary is a publication, usually a book, with a list of words in one or more languages, normally ordered alphabetically and explaining each word's meaning. There are several varieties of dictionaries. Depending on the purpose of a dictionary, the entries in the dictionary may contain various quantities of information about the word, such as its translation into another language, its etymology, a description of its meaning, its pronunciation, how it is used, etc. We define a dictionary as a comprehensive catalogue of words and information about these words. A dictionary can therefore be classified as a list.

Some dictionaries include cross-referencing between related entries. Such references are not regarded as part of the inherent structure of the dictionary; most dictionaries can therefore be classified as lists despite the inclusion of some non-hierarchical relations.

7.3 Index

The Latin word index originally meant indicator, and more specifically forefinger, the finger used for pointing at things (Merriam-Webster 2010c). Among others, Merriam-Webster (2010b) uses two definitions of the term index appropriate for listing in the context of KOSs. The first definition is a list (as of bibliographical information or citations to a body of literature) of some specified datum (as author, subject or keyword), usually arranged in alphabetical order. The second definition is a list of items (as topics or names) in a printed work that provides for each

item the page number where it may be found. An example of an index as defined in the first definition is an index card catalogue, commonly found in libraries before electronic indexing. An example of an index defined according to the second definition is the typical back-of-book index. Pepper (2002) uses the metaphor of a map to describe such an index — a concise and accurate map to the content of the book. We define an index as a specific type of catalogue, namely a catalogue of concepts and the attributes needed to uniquely identify those concepts along with one or more pointers to sources that are not part of the index, where more information about the concept can be found.

Most indices can be classified as lists. Indices, however, are often more complex in structure than mere lists. They may contain sub-entries and cross-references between entries. Sub-entries imply hierarchical relations between terms while cross-references entail other types of relations between the entries in an index. Such complex indices may be classified as thesauri.

7.4 Knowledge base

The need to develop computer applications to perform intelligent actions necessitated the design of innovative KOSs. Knowledge-base systems (KBSs) are designed to provide information sources equipped with machine-processible semantics that can be communicated between agents (Fensel et al. 2003). An agent in this context may be a software agent, or it may be a person.

Originally each KBS defined its own concept language to use as the basis of a query language for formulating questions about compound concepts in the knowledge base. The knowledge between different KBSs could therefore not be shared. Gruber's proposal (1993) that KBSs

should share their resources resulted in changes to the design of new KBSs (Swartout & Tate 1999). More assumptions used to be coded in the data in the form of inference rules, as opposed to being hard-coded in the software. As a result, communication between different KBSs was achieved through standardizing their data models. KBSs that are able to share knowledge are called ontologies in AI circles and are also likely to have data models that can be classified as an ontology when applying our classification of KOSs.

The data models of knowledge bases do not always comply with our definition of a KOS that can be classified as an ontology. They comply only if semantic inference rules are included in its data. Uschold (2003) maintains that most of the intelligence of contemporary web-applications is still hard-wired in the software. If this is the case, given that the data is likely to contain definitions of a variety of semantic relations, their data models can at best be classified as a thesaurus.

7.5 Lexicon

The word lexicon derives from the Greek word λεξικόν (lexicon) from λεξικός (lexikos, meaning of or for words). It is often used as a synonym for a dictionary, glossary or VR-thesaurus. Generally, the term refers to the utterances in a language, namely its vocabulary, including its words and expressions.

In AI applications dealing with natural language processing and machine translation, the term lexicon is used to refer to a machine-readable entity containing the lexical information required to support syntactic and morphological processing. It may therefore incorporate lexical elements

and language rules (Vanopstal et al. 2009). The lexical elements can either be in full forms or canonical base forms, while the language rules appear in the form of parts of speech facts, spelling and grammar rules, and morphological rules for creating new words. A lexicon can thus be classified as an ontology. It contains entities in the form of lexical elements, classified in terms of their usage along with language rules. A well-known example of a lexicon of the English language is WordNet (Fellbaum 1998; Miller et al. 1990). Legg (2007) observes that WordNet is often used as a formal ontology in spite of its simplicity.

7.6 Topic Map

The problem of how to merge indices gave rise to the development of the ISO/IEC 13250 (2003) standard. The application of this standard allows portability of indices and enables the creation of comprehensive indices to large collections such as the internet itself. The resulting structure is called a topic map.

Topic Maps is an ISO standard for describing knowledge structures and associating them with information resources. The topics, associations and occurrences that comprise topic maps allow topic map authors to describe complex structures.

```
drive(Person: driver, Vehicle: drivenBy) ~ pdv
involved-in(Person: actor, Accident: action) ~ pia
may-cause(pdv: action, BadWeather: condition, pia: effect)
```

Figure 7: Representation of the assertion in Figure 4 using LTM

Meta-relations can be created in topic maps through an action that is termed reification. Pepper (2010) explains that reification means turning a name, occurrence or association role (or even the topic map itself) into a topic in order to make assertions about the thing it represents. Figure 7 is a representation of the assertion shown in Figure 4 as a topic map using the Linear Topic Map Notation (LTM) (Garshol 2006).

Topic Maps Constraint Language (ISO 19756) (Moore & Bogachev 2010) (TMCL) provides means to specify cardinality as well as quantification restrictions for any item in a topic map. Furthermore, the use of a scope restriction in a topic map can be interpreted as being an inference rule in the form of a conditional statement; for example, if a given relation is scoped, the relation only applies in a specified context. Garshol (2008) experimented with bidirectional conversion between topic maps and ontologies that are expressed using Web Ontology Language (OWL) – the de facto standard markup language for web pages where the content needs to be processed by computer applications. Although he has identified constructs that cannot be converted in each of the directions, his practical experiments seem to illustrate that the expressiveness of TMCL exceeds that of OWL.

Our definition of an ontology requires the presence of inference rules, such as the ability to define meta-relations, restrictions such as quantification restrictions on attributes and cardinality restrictions on relations, as well as conditional rules. A topic map that makes use of these abilities can thus be classified as an ontology. There is, however, no consensus about whether topic maps should be classified as ontologies or not. Borgo (2004) explicitly states that topic maps are not ontologies.

8. Summary

The classification of KOSs presented here uses one of many facets that can be used to classify KOSs. The five broad classes of KOSs in our classification are called lists, taxonomies, lattices, thesauri and ontologies. We propose concise definitions for these terms as classes of our classification. These are not intended to replace established definitions of terms. They are presented to clarify the meaning of the terms and to advocate the proper usage of the terms. In particular we advise strongly against the free usage of the established CS term ontology to refer to a KOS that can be classified as a relationship list in Hodge's classification and which is classified as a thesaurus in our classification.

Other classification-related terms, namely catalogue, dictionary, index, knowledge base, lexicon and Topic Map, are described to clarify their meaning. These are all evaluated in terms of where they fit into our classification of KOSs.

We have uncovered the deplorable fact that many advances in CS have ignored the wealth of information that is embedded in existing classifications and the widespread usage of technologies that are inferior to standards and technologies that have been developed in LIS, in particular the ISO/IEC 13250 (2003) standard for topic maps which we deem ideal for formalizing ontologies.

In this paper we explicated reasons for the ostensible confusion caused by the inconsistent meanings that are attached to classification-related terms, most notably the term ontology. We propose a classification of KOSs that may partly eliminate this confusion. We hope that this will promote communication and collaboration within and across domains. A more comprehensive

understanding of work that has been done in the different domains will enable inter-disciplinary research to achieve higher levels of interoperability between systems that originated in opposing domains to eventually realize the Semantic Web dream.

References

ANSI/NISO Z39.19. 2005. *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*. <http://www.niso.org/standards/resources/Z39-19.html> (Retrieved March 14, 2014).

Aswani Kumar, Ch., Radvansky M. and Annapurna J. 2013. Analysis of a Vector Space Model, Latent Semantic Indexing and Formal Concept Analysis for Information Retrieval *Cybernetics and Information Technologies*, 12(1), 34 – 48.

Birkhoff, G. 1948. *Lattice theory* (revised edition). New York: American Mathematical Society.

Bo Ning, Z. 2013. Research on Web Information Retrieval based on Vector Space Model. *Journal of Networks*, 8(3), 688 - 695. doi:10.4304/jnw.8.3.688-695

Borgo, Stefano. 2004. *Classifying medical ontologies*.

<http://www.slideworld.org/viewslides.aspx/Classifying--Medical--Ontologies-ppt-86089>.

(Retrieved April 30, 2011)

Borgo, Stefano and Leitão, Paulo. 2007. Foundations for a core ontology of manufacturing. In R. Sharman, R. Kishore and R. Ramesh. eds., *Ontologies* (Vol. 14, p. 751-775). US: Springer.

Buitelaar P. and Cimiano P. eds. 2008. *Ontology Learning and Population: Bridging the Gap between Text and Knowledge* (Foreword). Amsterdam, The Netherlands: IOS Press.

Carpineto, C. and Romano, G. 2004. Exploiting the potential of concept lattices for information retrieval with credo. *Journal of Universal Computer Science*, 10 (8), 985 – 1013.

Chaffey, Dave, Ellis-Chadwick, Fiona, Johnston, Kevin and Mayer, Richard. 2009. *Internet marketing: Strategy, implementation and practice*. Edinburgh Gate, England: Pearson Education Limited.

Cimiano, Philipp, Hotho, Andreas, Stumme, Gerd and Tane, Julien. 2004. Conceptual knowledge processing with formal concept analysis and ontologies. In *Concept lattices* (Vol. 2961, pp. 199 – 216). Berlin / Heidelberg: Springer.

Corcho, Oscar and Gómez-Pérez, Asunción. 2000. A roadmap to ontology specification languages. In R. Dieng & O. Corby. eds., *Knowledge engineering and knowledge management methods, models, and tools* (Vol. 1937, pp. 80 – 96). Berlin / Heidelberg: Springer.

Dewey, M. (2011). Dewey decimal classification and relative index. 23rd edition; Mitchell J.S., Beall, J., Green, R., Martin G. and Panzer M.,eds., Dublin, Ohio: OCLC.

Fellbaum, Christiane.ed., 1998. *WordNet: An Electronic Lexical Database*. Cambridge, Mass: MIT Press.

Fensel, D., Hendler, J., Lieberman, H. and Wahlster, W..eds., 2003. Spinning the semantic web: Bringing the World Wide Web to its Full Potential. Cambridge, Mass: MIT Press.

Frank, A. U. 2006. Distinctions produce a taxonomic lattice: Are these the units of mentalese? In *Proceeding of the 2006 conference on formal ontology in information systems: Proceedings of the fourth international conference* (pp. 27–38). Amsterdam, The Netherlands: IOS Press.

Garshol, Lars Marius. 2004. Metadata? Thesauri? Taxonomies? Topic Maps! Making Sense of it all. *Journal of Information Science*, 30 (4), 378 – 391. Available from <http://jis.sagepub.com/cgi/content/abstract/30/4/378>

Garshol, Lars Marius. 2006. *The linear topic map notation definition and introduction* (Tech. Rep. No. Version 1.3 (Revision: 1.23)). Oslo: Ontopia A/S. Available from <http://www.ontopia.net/download/ltn.html> (Retrieved June 15, 2011)

Garshol, Lars Marius. 2008. TMCL and OWL. Maicher, Lutz. and Garshol, Lars Marius.eds., *Subject-centric computing. Fourth International Conference on Topic Maps Research and Applications, TMRA 2008, Leipzig, Germany, October 16-17, 2008, Revised Selected Papers.* (Leipziger Beiträge zur Informatik: XII) - ISBN 978-3-941152-05-2

Gilchrist, Alan. 2003. Thesauri, taxonomies and ontologies – an etymological note. *Journal of Documentation*, 59 (1), 7 – 18.

Grätzer, George A. 2011. *Lattice theory: foundation*. New York : Springer Basel.

Green, Rebecca. 2006. Vocabulary Alignment via Basic Level Concepts. *OCLC/ALISE research grant report*, OCLC Research. Available online at: <http://www.oclc.org/content/dam/research/grants/reports/green/rg2005.pdf?urlm=162177> (Retrieved March 14, 2014)

Green, Rebecca, Vizine-Goetz, Diane Zeng, Marcia Lea and Žumer, Maja. 2013. From modelling to visualization of topic relationships in classification schemes. *Classification and visualization: interfaces to knowledge: proceedings of the International UDC Seminar* (pp 177 – 190), 24-25 October 2013, The Hague, The Netherlands . Edited by Aida Slavic, Almila Akdag Salah and Sylvie Davies. Würzburg: Ergon Verlag, 2013.

Gruber, Thomas R. 1993. A Translation Approach to Portable Ontologies. *Knowledge Acquisition*, 5 (2), 199–220.

Hedden, Heather. 2010. *The Accidental Taxonomist*. Medford, NJ: Information Today Inc.

Hilera, J. R., Pagés, C., Martínez, J. J., Gutiérrez, J. A. and Marcos, L. de. 2011. An evolutive process to convert glossaries into ontologies. *Information Technology and Libraries*, 29 (4), 192–20.

Hjørland, Birger. 2007. Semantic and Knowledge Organization. *Annual Review of Information Science and Technology*, 41, 367 – 405.

Hodge, Gail. 2000. Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. *Digital Library Federation*, Council on Library and Information Resources, 1755 Massachusetts Ave., NW, Suite 500, Washington, DC 20036.

ISO/IEC 13250:2003 *Topic Maps* (2nd edition)

http://webstore.iec.ch/Webstore/webstore.nsf/ArtNum_PK/40946!opendocument.

(Retrieved January 10, 2014)

Jacob, Elin K. 2010. Proposal for a Classification of Classifications built on Beghtol's Distinction between "Naïve Classification" and "Professional Classification". *Knowledge Organization*, 37(2), 111 -120.

Kang, Jang-Mook, Im, Yong Soon, Lee, Ki Young, Lim, Myung Jae, Lee, Young Dae, Kang, Eun-young, Kang, Myeon-Koo and Oh SunJin. 2009. A Study for Semantics Participation Platform Architecture Using RDF/OWL. In *Proceedings of the 2009 International Conference on Hybrid Information Technology* (pp. 512–515). New York, NY, USA: ACM. Available from <http://dl.acm.org/citation.cfm?doid=1644993.1645088> (Retrieved March 12, 2014)

Kempf, Andreas Oskar, Ritze, Dominique, Eckert, Kai and Zapilko, Benjamin. 2014. New Ways of Mapping Knowledge Organization Systems: Using a Semi-Automatic Matching Procedure for Building up Vocabulary Crosswalks *Knowledge Organization*, 41(1), 66 – 75.

Klyne, Graham and Carroll, Jeremy J. 2004. Resource description framework (RDF): Concepts and abstract syntax – W3C recommendation. Available <http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/> (Retrieved March 12, 2014)

Lambe, Patrick. 2007. *Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness*. Oxford, UK: Chandos Publishing Limited.

Laukaitis, Algirdas and Vasilecas, Olegas. 2007. Formal concept analysis and information systems modeling. In *Proceedings of the 2007 international conference on computer systems and technologies* (pp. 17:1–17:6). New York, NY, USA: ACM.

Legg, Catherine. 2007. Ontologies on the Semantic Web. *Annual Review of Information Science and Technology*, 41 (1), 407–451. Available from DOI: 10.1002/aris.2007.1440410116 (Retrieved March 12, 2014)

McComb, D. 2003. Terms: Vocabulary, Taxonomy, and Ontology. In *Semantics in Business Systems* (pp. 41 – 64). Burlington: Morgan Kaufmann.

McGuinness, Deborah L. and van Harmelen, Frank. 2004. OWL web ontology language – W3C recommendation. Available <http://www.w3.org/TR/owl-features/> (Retrieved March 12, 2014)

Merriam-Webster. 2010a. *catalog* — *Merriam-Webster Online Dictionary*.

<http://www.merriam-webster.com/dictionary/catalog>. (Retrieved December 17, 2010)

Merriam-Webster. 2010b. *index* — *Merriam-Webster Online Dictionary*.

<http://www.merriam-webster.com/dictionary/index>. (Retrieved December 17, 2010)

Merriam-Webster. 2010c. *index* — *Word-Origins.com, WebFinance Inc.* <http://www.word-origins.com/definition/index.html>. (Retrieved December 17, 2010)

Miles, Alister and Bechhofer, Sean. 2009. SKOS simple knowledge organization system reference - W3C recommendation. Available <http://www.w3.org/TR/skos-reference/> (Retrieved March 12, 2014)

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. J. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3 (4), 235-244.

Moore, G. & Bogachev, D. 2010. *ISO 19756: Topic Maps Constraint Language*. Available from <http://www.isotopicmaps.org/tmcl/> (Retrieved March12, 2014)

Nguyen, Phillip H.P., Kaneiwa, Ken and Nguyen, Minh-Quang. 2010. Ontology inferencing rules and operations in conceptual structure theory. In T. Meyer, M. Orgun and K. Taylor. eds., *Conferences in research and practice in information technology (crpit)* (Vol. 122). Adelaide, Australia: Australian Computer Society, Inc.

Nguyen, Phillip H.P., Kaneiwa, Ken, Corbett, Dan R. and Nguyen, Minh-Quang. 2009. Meta-relation and ontology closure in conceptual structure theory. *Artificial Intelligence and Law*, 17, 291-320. Available from <http://dx.doi.org/10.1007/s10506-009-9082-z> (Retrieved March12, 2014)

Noy, N. F. and McGuinness, D. L. 2001. *Ontology development 101: A guide to creating your first ontology*, Palo Alto, California: Stanford Knowledge Systems Laboratory. Available from http://protege.stanford.edu/publications/ontology_development/ontology101.pdf (Retrieved March12, 2014)

OED. 1989. *Taxonomy — The Oxford English Dictionary*, OED Online 2nd edition. Available from <http://0-dictionary.oed.com.innopac.up.ac.za/cgi/entry/50247829> (Retrieved August 15, 2008)

OWL 2 Web Ontology Language: structural specification and functional-style syntax (2009).
W3C Recommendation 27 October 2009. Edited by Motik Boris, Patel-Schneider Peter F
and Parsia Bijan. Available at: <http://www.w3.org/TR/2009/REC-owl2-syntax-20091027/>
(Retrieved March 12, 2014)

Pepper, Steve. 2002. *Ten Theses on Topic Maps and RDF*. Available at:
<http://www.ontopia.net/topicmaps/materials/rdf.html>. (Retrieved March 12, 2014)

Pepper, Steve. 2010. Topic Maps. In Bates Marcia J. and Maack Mary Niles. eds.,
Encyclopedia of Library and Information Sciences (Third edition., pp. 5247 – 5259).
Abingdon, Oxford: Taylor & Francis.

Priss, U. 2000. Lattice-based information retrieval. *Knowledge Organization*, 27 (3), 132 –
142.

Priss, U. 2006. Formal concept analysis in information science, *Annual Review of
Information Science and Technology*, 40(1):521 - 543.

Roget, Peter Mark. 1912. *Thesaurus of English Words and Phrases*. London: Dent.

Sakharov, Alex. n.d. *First-Order Logic*. From *MathWorld* — A Wolfram Web Resource, created by Eric W. Weisstein. <http://mathworld.wolfram.com/First-OrderLogic.html>. (Retrieved May 10, 2011)

Sowa, J. F. 1984. *Conceptual Structures: Information Processing in Mind and Machine*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Spärck-Jones, Karen. 1992. Thesaurus. In Shapiro S. C. ed., *Encyclopedia of Artificial Intelligence* 1 (2), 1605 – 1613. New York, NY, USA: John Wiley & Sons Inc.

Stumme, Gerd. 2002. Formal Concept Analysis on its Way from Mathematics to Computer Science. Proceedings of the 10th International Conference on Conceptual Structures. *Lecture Notes in Computer Science*, Vol. 2393, pp. 2–19. Berlin / Heidelberg: Springer.

Swartout, W. and Tate, A. 1999. Ontologies. *IEEE Intelligent Systems and their Applications*, 14 (1), 18 – 19.

Tudhope, D. and Binding, C. 2008. Faceted Thesauri. *Axiomathes*, 18, 211– 222.

Tur, Gohkun and De Mori, Renato. 2011. *Spoken language understanding: Systems for extracting semantic information*. New York, NY, USA: John Wiley & Sons Inc.

Uschold, M. 2003. Where are the semantics in the semantic web? *AI Magazine*, 24, 25 – 36.

Valtchev, P., Grosser, D., Roume, C. and Hacene, M. R. 2003. Galicia: An Open Platform for Lattices. In *Using Conceptual Structures: Contributions to the 11th International Conference on Conceptual Structures* (pp. 241–254). Aachen, Germany: Shaker Verlag.

Vanopstal, Klaar, Vander Stichele, Robert, Laureys, Godelieve and Buyschaert, Joost. 2009. Vocabularies and retrieval tools in biomedicine: Disentangling the terminological knot. *Journal of Medical Systems*, 1-17. Available from <http://dx.doi.org/10.1007/s10916-009-9389-z>

Wikipedia. 2008a. *Systema Naturae* — *Wikipedia, The Free Encyclopedia*.

[http://en.wikipedia.org/w/index.php?title=Systema Naturae&oldid=241992965](http://en.wikipedia.org/w/index.php?title=Systema_Naturae&oldid=241992965).

(Retrieved October 8, 2008)

Wikipedia. 2008b. *Taxonomy* — *Wikipedia, The Free Encyclopedia*. <http://en>

[.wikipedia.org/w/index.php?title=Taxonomy&oldid=231989736](http://en.wikipedia.org/w/index.php?title=Taxonomy&oldid=231989736). (Retrieved August 18, 2008)

Wikipedia. 2010. *Ontology* — *Wikipedia, The Free Encyclopedia*.

<http://en.wikipedia.org/w/index.php?title=Ontology&oldid=377803633>. (Retrieved August 10, 2010)

Wikipedia. 2011. *Thesaurus* — *Wikipedia, The Free Encyclopedia*.

<http://en.wikipedia.org/w/index.php?title=Thesaurus&oldid=411745255>. (Retrieved February 8, 2011)

Wiktionary. 2010a. *Dictionary* — *Wiktionary, A wiki based open content dictionary*.

<http://en.wiktionary.org/w/index.php?title=dictionary&oldid=10921188>. (Retrieved November 17, 2010)

Wiktionary. 2010b. *Thesaurus* — *Wiktionary, A wiki based open content dictionary*.

<http://en.wiktionary.org/w/index.php?title=thesaurus&oldid=11805174>. (Retrieved November 17, 2010)

Wille, Rudolf. 1982. Restructuring lattice theory: an approach based on hierarchies of concepts. In Rival I., ed., *Ordered sets* (pp. 445–470). Dordrecht-Boston: Reidel.