"Thesaurus" and "Ontology:" A Study of The Definitions Found in the Computer and Information Science Literature, by Means of an Analytical-Synthetic Method

Alexandra Moreira,* Lídia Alvarenga,** and Alcione de Paiva Oliveira***

*Grupo de Sistemas de Apoio à Decisão, Universidade Federal de Viçosa, Av. P.H. Rolfs S/N,Viçosa MG, Brazil 36570-000, <xandramoreira@yahoo.com.br>

**Escola de Ciência da Informação-Universidade Federal de Minas Gerais, Avenida Antônio Carlos 6627, Pampulha, Belo Horizonte MG, Brazil 31270-901 <lidiaalvarenga@eci.ufmg.br>

***Departamento de Informática, Universidade Federal de Viçosa, Av. P.H. Rolfs S/N, Viçosa MG, Brazil 36570-000 <alcionepaiva@yahoo.com>

Alexandra Moreira is a member of the ontology group at the Federal University of Viçosa, Minas Gerais, Brazil. She holds a Masters degree in Information Science from the Federal University of Minas Gerais, Brazil. She teaches courses in specialized information sources and information treatment. Her research interests include concept theory, ontology construction, knowledge representation and knowledge organization.

Lidia Alvarenga is a senior professor at the Information Science School of Federal University of Minas Gerais State, in Belo Horizonte, Brazil. Her current research interests include knowledge organization in digital libraries and bibliometric studies for epistemological domains, dialoging with Michel Foucault´s Knowledge Archaeology theory. She is coordinator of an institutional research group in digital libraries, which has among its goals to improve studies on description of digital objects and faceted categories in the context of web semantics and hypertext arquitectures.

Alcione de Paiva Oliveira is head of Computer Science at Federal University of Viçosa, Minas Gerais, Brazil. He holds a Doctor degree in Computer Science from the Catholic University of Rio de Janeiro, Brazil. He teaches courses in artificial intelligence and computer science theory. His research interests include multi-agent systems, knowledge representation and knowledge organization.

Moreira, Alexandra, Lídia Alvarenga, and Alcione de Paiva Oliveira. **"Thesaurus" and "Ontology:" A Study of the Definitions Found in the Computer and Information Science Literature, by Means of an Analytical Synthetic Method**. *Knowledge Organization, 31(4).* 231-244. 44 refs.

ABSTRACT: This is a comparative analysis of the term ontology, used in the computer science domain, with the term thesaurus, used in the information science domain. The aim of the study is to establish the main convergence points of these two knowledge representation instruments and to point out their differences. In order to fulfill this goal an analytical-synthetic method was applied to extract the meaning underlying each of the selected definitions of the instruments. The definitions were obtained from texts well accepted by the research community from both areas. The definitions

were applied to a KWIC system in order to rotate the terms that were examined qualitatively and quantitatively. We concluded that thesauri and ontologies operate at the same knowledge level, the epistemological level, in spite of different origins and purposes.

1. Introduction

The motivation to develop a comparative study of the meaning of "thesaurus" and "ontology" in information and computer science areas relies on the observation of a lack of understanding about those representation and information recovery instruments, which might lead to serious problems in discourse. Nowadays, there is a great demand for development of systems that work with retrieval, sharing and exchange of information. To support those systems, new knowledge organization instruments appear each day, which have been called "ontologies." The use of the term "ontology" to denote a structure of terms and the relations between them in one specific domain is more common in the computer science area, more specifically, in the artificial intelligence domain. Some researchers (Jasper & Uschold 1999, Fensel 2000 and 2001) consider thesauri to be simple ontologies, while a complex ontology, according to these authors, demands more relations than those traditionally presented in a thesaurus. As an example we can mention the following passage of Fensel et al. (2001, 38): "Large ontologies such as WordNet provide a thesaurus for over 100,000 terms explained in natural language." In this sense, a thesaurus can be understood as a type of ontology directed towards the organization of terms. Such statements denote a blurry frontier between the concepts of thesaurus and ontology that demands further clarification. Would thesaurus and ontology be terms that denote the same instrument or would they denote different objects and therefore, indistinct use of the terminology would be inappropriate? In the worst case, a terminological mistake would be established and the consequences of this confusion would be the lack of agreement on what characterizes each tool, resulting in possible inadequate use. The development of the tools is especially harmed, once the terminological confusion hinders information exchange among researchers. The use of the term ontology became very popular in the computer science field, mainly in the sub-area of knowledge representation. One of the main objectives of the use of ontologies in computer science is to allow the construction of interoperable knowledge bases. Under the ontology designation, tools have been created to help document storage and recovery in computational systems (Guarino, Masdolo and Vetere 1999), information extraction in natural language texts and in e-commerce systems, exchange of information among intelligent agents (Cranefield, Purvis and Nowostawski 2000), automatic knowledge acquisition tools (Duarte 2002), system modeling (Villela, Oliveira, and Braga 2004), and several other tasks related to the use and the representation of knowledge.

The justification for a study to understand the term "ontology" in information science arises from the fact that knowledge representation is also an object of study of the area. Therefore the contribution of ontologies for information science must be investigated. Moreover, concept content analyses study is one attribute of information science, especially when the analytical-synthetic method is applied, which is a classic instrument of its theoretical body. The comparison of the term "thesaurus" aims to investigate why some computer science researchers define it as an "informal ontology."

This work presents a comparative study based on the most frequent definitions, aiming to contribute terminological clarification and understanding of the two concepts. Analytical-synthetic method (Dahlberg 1978) is applied to definitions found in the literature for both terms in order to perform a content analysis of the term. The paper is organized as follows. The next section presents origins and some definitions of "thesaurus;" in section 3 origins and some definitions of ontology are presented; in section 4 the methodology is defined; in section 5 empirical results and their interpretation are shown, finally, in section 6 the conclusions are presented.

2. Thesaurus

The term "thesaurus" has its roots in the Greek and Latin languages and it means "treasure." This term became popular with the publication of Peter Mark Roget's analogical dictionary, in London, in 1852, entitled *Thesaurus Of English Words And Phrases*. Roget called his dictionary "thesaurus" – at one time the term also designated vocabulary, dictionary, or lexicon. The difference was that Roget's dictionary was a vocabulary organized by meaning and not by alphabetical order. Roget's thesaurus established a common denominator for vocabularies that relate

their terms using a certain type of semantic relationship.

It is important to emphasize Mortimer Taube's contribution from his Uniterm system in 1951 that, according to Lancaster (1986), can be considered responsible for the appearance of the thesaurus. Uniterm was composed by a set of records, where each record contained a single word and the numbers of the documents associated with each word. According to Gomes (2004), Uniterm was based on the hypothesis that each idea could be represented by a single word. The evolution of Uniterm resulted in the creation of the first thesaurus, developed by the Information Engineering Center of DuPont in 1959.

In the sixties, Vickery presented four meanings for the term thesaurus in the information science literature, and the most accepted meaning was an alphabetical list of words, where each word is followed by a list of related words (Foskett 1985, 270). Howerton (Currás 1995, 85) defines thesaurus as "an authority file, which can lead the user from one concept to another via various heuristic or intuitive paths."

Another work about thesauri that should be cited is that of the program Unisist (UNESCO, 1973, 6) that defines the term "thesaurus" for the information science area under two aspects: the structural and the functional. In the first case, it is "a dynamic controlled vocabulary of terms related semantically and by generic relation covering a specific knowledge domain." In the second view it is "a terminological control device used in the translation of the natural language of the documents, from the indexers or from the users in a more restricted system language (documentation language, information language)."

From the earliest initiatives, such as Roget's, until the present, the thesaurus evolved its definition and its theoretical and methodological construction, by the introduction of new cognitive models and user centered approaches. A current definition, resulting from this evolution is the one from Currás (1995) that states:

> Thesaurus is a specialized, normalized, post-coordinate language used for documentaries means, where the linguistic elements that composes it – single or composed terms – are related among themselves syntactically and semantically.

In other words, "specialized language" is understood to be one that is used in a restricted domain. "Normalized," is understood to be a controlled language where the linguistics units are terms and, finally, "post-coordinate language" is understood to be the terms combined at the time of use, in opposition to the pre-coordinated indexing languages whose terms designate complex subjects that are coordinated prior to use (a subject heading list, for instance).

Another definition was established by the National Information Standards Organization, in a document that sets the *Guidelines for the Construction, Format, and Management of Monolingual Thesaurus* (ANSI/NISO Z39-19-1993), where thesaurus is defined as:

> A controlled vocabulary arranged in a known order and structured so that equivalence, homographic, hierarchical, and associative relationships among terms are displayed clearly and identified by standardized relationship indicators that are employed reciprocally.

2.1 Lines of Evolution

For the study of the genesis and the evolutionary line of the thesaurus, it is worth mentioning the work done by Lancaster (1986). The author emphasizes the difficulty of showing this evolution accurately, once the influence chain is not clear. The historical evolution of the thesaurus is divided in two lines: one that has its base in the Uniterm system, introduced by Mortimer Taube in 1951; and the other, influenced by the Faceted Classification Theory. The first of the two lines comes from the alphabetical approach of North America, more specifically the United States, and on the other line it comes from the bibliographical classification of Europe, more particularly the United Kingdom, influenced by the work of Ranganathan in 1930. The Ranganathan analytical-synthetic study (faceted) established solid bases for classification methods, affecting subject alphabetical indexing and thesaurus construction (Thesaurofacet). Subject alphabetical indexing starts in the United States with Charles Ammi Cutter, in his *Rules for a Dictionary Catalogue* published in 1876. Both lines of thesaurus evolution tend to converge in the standard ISO 2788, in the preliminary edition of the second edition of 1983 and by the standard BS 5723 (British Standards).

The two evolutionary lines, European and American, possess some important distinctions due to the differing motivations for their development. The American line adopts a more pragmatic evolution,

motivated by the necessity of improvement of the limitations of a previous documentary language, the Uniterm.

On the other hand, the European line, mainly the line of Thesaurofacet, as it is based in Ranganathan's classification theory (Lancaster, 1986, p.33), applies the use of categories for organization of the concepts in a domain. The use of categories to frame the concepts allows a better organization of the hierarchies and a more appropriate positioning of the terms associated with the concepts.

Another more recent line of evolution, whose mention is indispensable, is that of the concept-based-thesaurus, also known as *terminological tesauros*. According to Campos (2001), this type of thesaurus arose from the junction of Concept Theory (developed by Dahlberg starting in the decade of the 1970s) with Classification Theory. Among the contributions of Concept Theory is a better understanding of the concept and of the term, the organization of concepts through categories, and the use of definitions for the positioning of a concept in the system.

The discussion that we present about thesaurus and ontology is independent of the line of evolution of the thesaurus. However, the distinctions we present will be more evident in the case of the thesaurus that uses categories to organize its elements.

2.2 Thesaurus Function

According to Currás (1995 translation ours), the thesaurus was adopted: "in the documentation area, associated with the form of organization of the indexing/recovering vocabulary." A thesaurus can work in an organizational environment for the representation of documents' subjects, as well as in informational searches. Representation of the subjects of documents is accomplished by the indexer, who analyzes the document, identifies its contents, and "translates" the contents into terms allowed by the thesaurus.

A thesaurus can be used not only to aid the elaboration of queries accomplished by the user, in informational searches, but also for the indexer during the classification process. For the two types of user, the thesaurus, by means of its structure of terms and relationships, helps to find the best term or terms that denote a subject. Therefore, the thesaurus is a very important component in a recovery system as it accomplishes the tasks of: determining which term can be used in the system; determining which term can be used in the search to have a satisfactory result; and allowing the introduction of new term in its structure of terms in order to bring user language closer to system language and to accomplish changes of senses of the existing terms.

3. Ontology

Ontology is one division of philosophy. According to García Morente (1964), philosophy has two great divisions: ontology (the being's theory), related to the philosophy of the antique and of the medium age; and gnosiology (theory of knowledge), related to the modern age. Ontology, in general, can be understood as the being's theory (*ontos* for being and *logos* for word). It is the theory of the objects, of the ontic structures, and of what exists.

Although the study of what exists can be traced back to Aristotle and Plato, the use of the term ontology to designate this branch of philosophy is much more recent, having been introduced around the 17th and 18th centuries by German philosophers.

According to Welty (2001) the term was coined in 1613 by Rudolf Goclenius and, apparently, in an independent way by Jacob Lorhard. The term ontology is mentioned briefly by Goclenius on page 16 of the *Lexicon philosophicum, quo tanquam clave philosophiae fores aperiuntur, Informatum opera studio Rodolphi Goclenii* where it says "ontologia, philosophia de ente" (Mora1963). But, as pointed out by Mora, Christian Wolff was responsible for making the word ontology popular in philosophical circles: "The word appears in the title of his *Philosophia prima sive ontologìa methodo scientifica pertractata, qua omnes cognitionis humanae principia continentur,* published in 1730."

3.1 Ontology in Computer Science and Artificial Intelligence

The term ontology began to be used in Computer science, in the sub-area of artificial intelligence (AI), in the early nineties, in projects whose goal was to organize big knowledge bases, like CYC (Lenat & Guha, 1990) and Ontolingua (Gruber, 1992). From the 1970s (Russel & Norvig, 1995) AI researchers had been concerned with the organization and manipulation of knowledge bases, but starting in the 1990s there was a urge for creation of knowledge bases that could be shared and reused. This urge was due to the perception that the complex problems should be attacked by different systems as a network of multi-agents acting in a cooperative/competitive

base. In addition, the effort for creation of a knowledge base can be very expensive, and the reuse and sharing of the bases can reduce the costs. However, the sharing of knowledge bases can only happen if there is a clear understanding of the "ontological commitments" associated with the bases. Ontological commitments are understood as the choices that were made to select a certain group of concepts instead of others (Valente, 1995, p.34). In other words, the ontological commitments determine what is relevant in a certain domain so that it is worth being represented in a knowledge base. For instance, if one chooses to represent the object "book" through a predicate in a logical language, he is committed to the existence of this property in the domain. In other words, that object that possesses the property of being a book exists. Even the choice of the representation language reveals some ontological commitments. For instance, the use of the first order logic reveals the ontological commitments to the existence of facts, objects and relationships (Russell 1995, 166). The sharing of ontological commitments makes it possible to communicate between agents (human or not), to establish common comprehension of a domain.

The explicit and formal registration of the ontological commitments is what, in most cases, has been called an ontology in the scope of artificial intelligence. While the traditional knowledge bases accumulated the knowledge necessary to assist a specific application, an ontology should have the property of being used in several applications and in distributed applications, as in multi-agent systems, supplying necessary support for information exchange among agents. Among the problems that can benefit from the use of ontologies we can mention: knowledge representation, knowledge reuse, knowledge sharing, knowledge acquisition and knowledge integration; natural language processing; automatic translation; information exchange among systems, agents, companies or people.

However, there still is no consensus about the interpretation of the term. Even though there exists an agreement that ontology is for AI a registration of the ontological commitments, the form in which the registration is made is still the subject of debate. Among researchers that tried to clarify the interpretation of the term in artificial intelligence we can mention Guarino and Giaretta (1995) and Poli (2001). Guarino and Giaretta present seven interpretations for the term ontology that subsist in artificial intelligence and knowledge representation. Poli analyzes some of the current definitions of the term in the AI field under orientation criteria (object oriented and concept oriented ontologies) and under domain independence criteria. According to him, in the correct sense of the word, ontologies should be object oriented and domain independent, while concept oriented and domain dependent ontologies should be the most spurious. Some of the most common interpretations for the term ontology, according to the obtained definitions are:

1. Ontology as an informal conceptual system, which underlies a particular knowledge base.
2. Ontology as a representation of a conceptual system via a logical theory.
3. Ontology as the vocabulary used by a logical theory.
4. Ontology as a specification of a conceptualization.

These interpretations are a subset of the interpretations presented by Guarino and Giaretta. The first considers ontology as a conceptual system that can underlie a knowledge base. In this case ontology belongs to the conceptual level and not to the symbolic level. In interpretations 2-4 the term ontology is interpreted as denoting an entity at the symbolic level. The second interpretation defines ontology as a special type of knowledge base (the term "knowledge base" is used in the sense of a group of sentences describing the state of a domain in the form of a logical theory), differing from a common knowledge base in the sense that it possesses a special type of knowledge (knowledge independent of a particular domain state of affairs) that serves a specific purpose (communication, queries, etc.). According to interpretation 3, ontology is just the vocabulary used in a logical theory, and the level of formalization of this vocabulary can vary from one ontology to another. Interpretation 4 establishes ontology as a specification of a conceptualization, and a conceptualization can be understood as a set of ontological commitments. This last interpretation is the one that is gathering the largest number of followers in the AI community.

4. Methodology

This is research that aims for a terminological explanation and characterization of two tools used in the organization of the knowledge: thesaurus and ontology. All of the theoretical and material data used

were collected from bibliographical sources related to the studied domains, including philosophy, information science, and computer science. In the case of computer science, we looked more specifically in the sub-area of artificial intelligence. The material used were the definitions collected from bibliographical sources related to the computer science and information science areas. Regarding philosophy, the bibliographical sources were used only for theoretical data, whose purposes were to supply elements for the qualitative analysis.

Definitions were studied through the analytical-synthetic method. The techniques adopted were quantitative and qualitative content analysis. The use of qualitative analysis together with the quantitative enriches the analysis of the data. The quantitative aspects are the starting points and serve as support for the analysis, and the qualitative aspects supply a better understanding of the registered data.

The technique proposed here can also be seen as a technique of content analysis. Content analysis can be understood as "a research technique for objective, systematic and quantitative description of the communication of explicit content" (Berelson apud Marconi & Lakatos 1982, 99). For Ander-Egg (1978, 178) it is "a well known technique to investigate the mass communication content by categorization of the communication elements." In this work the definitions were analyzed using systematic categories, which emerged from the definitions themselves, producing the quantitative results.

4.1 Analytical-synthetic method

The method proposed by Dahlberg (1978) for the analysis and structuring of concepts has its roots in the analysis of true propositions of a concept, generating a hierarchy of characteristics, yielding one category as the most generic characteristic, and synthesizing these characteristics in the form of a term or a name whose meaning is established precisely through a definition. Because it includes as much of an analysis stage as a synthesis stage, the method is termed analytical-synthetic. This method provides a safe way for understanding the intension of a concept for insertion in the structure of concepts of a specific domain.

To accomplish the general objective of determining the meaning of the terms "ontology" and "thesaurus" in the described areas it is necessary to apply a method that allows the emergence of the underlying meanings in each definition as well as subsequent comparison. For this reason, the method chosen for this task is the analytical-synthetic method. The analytical-synthetic method has been applied in the scope of information science in situations that involve the construction of conceptual structures, as in thesaurus construction (Gomes 1990), or in situations that involve the comprehension and appropriate definition of a concept, as in Alvarenga (1993).

Alvarenga used the analytical-synthetic method to analyze and to propose a definition for the concept of "official publications." Due to similarity with the present work, the steps used by Alvarenga in the development of her work were adapted for this study. We also included the use of a KWIC tool to separate the terms occurring in the definitions. The methodological steps used in this research were the following:

- From the literature, the definitions and concepts about ontology in computer science and about thesaurus in information science were extracted.
- The definitions were applied to a program that produces an alphabetical ordering of words in agreement with the KWIC indexation system. KWIC (Keywords in context) it is a form of text analysis where the words are listed alphabetically, maintaining the whole sentence and preserving the context of the occurrence of the word. Prior to this phase stop-words were eliminated. The output of the KWIC program also showed the total number of occurrences of the word.
- Gathering of the terms in categories or wide classes. The categories were obtained from the similarities among the attributes.
- Discussion of categories. In this stage the chosen categories were analyzed as well as the attributes and characteristics related to them.
- Transcription into an occurrence array of the concepts in each category. This array is fundamental for comparative analysis of the proposed content.
- Analysis of the definitions and concepts according to their occurrence in certain categories. In this way, it is possible to verify differences and similarities between two concepts.
- A wide discussion of the results in light of prior knowledge formed the qualitative approach that led to the conclusions.

4.2 The text selection process

Due to the great number of documents it was not possible to analyze the entire bibliography related to

the themes. Therefore, it was necessary to establish criteria for document selection. We opted to select the most relevant texts found in the literature. In the case of the thesaurus, as it is an exhaustively discussed instrument in information science, the term is very consolidated, so we looked up the fundamentals texts about the subject such as Lancaster (1986 and 1987) and Gomes (1990). For the term "thesaurus" we found six definitions, two of them in the same document (UNESCO). This small number of definitions can be attributed to the maturity that the term has reached in the area, so that is not necessary to define the term very frequently anymore.

On the other hand, the term "ontology" in computer science is not well established and it possesses several definitions and senses. So the criterion adopted was to select the texts based on the number of citations to them according to the queries made to the CiteSeer website (Lawrence *et al*. 1999). CiteSeer (also known as ResearchIndex) is a digital library of scientific literature, whose goal is to aid the dissemination of scientific literature related mainly to computer science. CiteSeer analyzes, through an autonomous computational system, documents in electronic format, removing citations and storing them in an index for subsequent consultations. The system operates without any human intervention.

In a query in CiteSeer using the word "ontologies," documents that totaled 2674 citations were returned. The articles were ordered by a function that takes into account the number of citations and the date of publication of the article, where the most recent articles received a greater weight. Figure 1 shows a segment of the result of the query containing the three best-classified articles.

The list returned from CiteSeer revealed that beyond a given point, 196 of the list were articles with just one citation, demonstrating a concentration of citations in few articles. Some of the articles more frequently mentioned belonged to a group of few authors. An increasing number of citations began in the decade of the 1990s, showing that it was at that time that the interest for the subject emerged in the computer science. For the term "ontology," thirteen definitions were used, extracted from twenty-nine articles returned from the query accomplished in CiteSeer.

4.3 Term selection

The selection of terms obeyed the onomasiological approach. In other words, terms selected denoted concepts of the domain. As a consequence of this approach, some composed terms were not dismembered into constituent terms, once they denoted relevant concepts in the domain. On the other hand, other composed terms were dismembered into their constituent terms, as they denoted existing concepts in the context. For instance, the term "knowledge base" was not dismembered into "base" and "knowledge," as the composed term denotes an entity known by the artificial intelligence community.

To obtain the terms it was necessary to use the output of the KWIC generating program to assure that each element formed a semantic unit, once the program generated an output of words and not of terms. This was done by taking each word emitted by the program, and verifying whether it denoted a concept of the domain. Otherwise, it was designated a composition with the words in its neighborhood and a new verification of the meaning was accomplished.

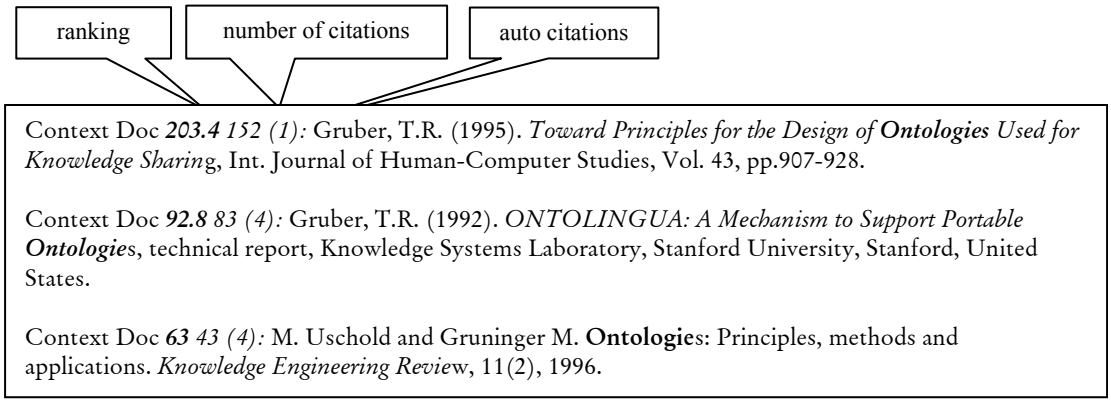

Figure 1. - *Segment of the result of the query using the word "ontologies"*

5. Results

The application of the methodology resulted in the elaboration of categories that was used in the quantitative and in the qualitative analysis presented in the following discussion.

5.1 Categories

After the selection of the definitions and extraction of the terms, the elaboration of the categories took place. The objective was to join, in the same group, similar concepts denoted by different terms. The categories were established through the similarities among the inherent characteristics of the terms in an inductive process. It should be considered that the definitions were taken from different areas, with their own terminologies and, therefore, in certain situations, in spite of the use of different terms, they may be denoting the same concept. The use of categories seeks to balance the terms, once the terms sheltered under the same category denote related meanings. The categories were applied to both contexts: ontologies and thesaurus. Table 1 shows the list of categories obtained.

In the "object" category there was no term that occurred in both definition sets. That was also the case of the "processes" category, probably due to the difference of activities between the two areas. In the "language" category the terms "natural language" and "terms" came from both contexts. Some terms are related, as is the case of "vocabulary" from the ontologies context and "controlled vocabulary," from the thesaurus context. In the "knowledge space" category the term "domain" is used in both contexts. In the "semantic content" category the terms "concept" and "knowledge" are used in both contexts and the terms "meaning" (ontology) and "semantics" (thesaurus) possess the same sense. Although the "attribute" category sheltered many terms, just the terms "hierarchical/hierarchically" and "constraint/constrain/restrict" happened in both sets of definitions. In the "systematization" category the term "relation/relations/related" is used in both contexts. The category "agents" didn't register any term originated from the ontologies context.

Category	*Description*
Object	Gathers the terms related to the concrete objects mentioned in the definitions, such as "knowledge bases" and "documents".
knowledge space	Gathers the terms related to the knowledge space delimitation, such as "subject" and "domain".
Semantic Content	Gathers the terms associated with elements of meaning, such as "concepts" and definitions".
Language	gathers the terms related to the language type used by the objects that are being defined, such as "natural language", "controlled vocabulary" and "vocabulary".
Process	Gathers the terms related to the process mentioned in the definitions, such as "indexing", "description" and "restriction".
Attribute	Gathers the terms associated with the properties of the objects that are being defined, such as "normalized", "formal" and "explicit".
Systematization	Gathers the terms related to the organization and structuring, such as "system", "scheme" and "structure".
Agents	Gathers the terms related to agents, such as "users".

Table 1. *List of categories.*

Table 2 presents the terms' occurrence quantification for each category. The first column of the table lists the categories. The second and the third columns register the number of terms from the ontologies definitions that occur in the category in absolute numbers and in percentages, respectively. The fourth and the fifth column register the number of terms

Categories	*Ontology*		*Thesaurus*	
Object	3	5%	3	6%
Language	12	18%	17	35%
knowledge space	8	12%	1	2%
Semantic Content	7	11%	4	8%
Process	10	15%	3	6%
Attribute	14	22%	12	25%
Systematization	11	17%	6	12%
Agents	0	0%	3	6%

Table 2. *Frequency of Occurrence of the terms for each category*.

from the thesaurus definitions that occur in the category in absolute numbers and in percentages, respectively. All of the occurrences of repeated terms in different definitions are counted.

that this concept is an important element in the ontologies identification. Figure 2 displays the number of occurrences of terms and the number of categories pointed by each definition.

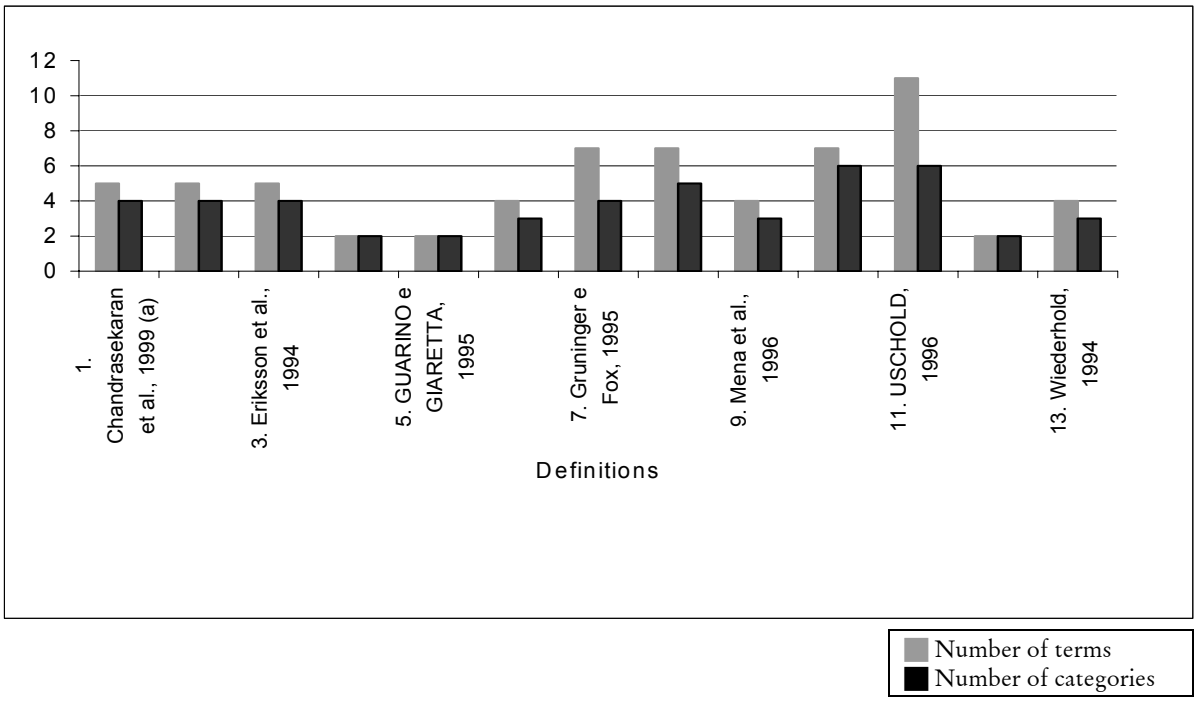

Figure 2. *Occurrence of terms and categories for each definition.*

There is a concentration of terms for thesaurus definitions in the categories language and attribute. With ontology definitions there is a better distribution, with a bigger emphasis in the attribute, systematization, and language categories. The category "agents" didn't receive any term originating from the definitions of ontology. This absence can be attributed to a greater concern with storage and processing by computational devices than with manipulation by users. It is also noticed that the category "process" registered a greater frequency of occurrence of terms originating from ontology than from the thesaurus definitions.

The terms "domain," "term," and "related/relations" were those that occurred most frequently in the ontologies definitions. The occurrence of the first indicates that ontology in computer science is related to specific domains and not to domain-independent knowledge, as is the case of ontology in philosophy. The frequency of occurrence of the term "term" indicates that in many cases ontology in computer science is an attempt to establish terminology for a certain domain. The frequency of occurrence of the term "related/relations" indicates

In spite of a smaller number of definitions about thesaurus, they have come up with almost the same number of terms that emerged from ontology definitions. The terms that happen more frequently are the terms "related/relation" and "term." The occurrence of these terms shows that the thesaurus focuses the registration of terms and their relationships and any definition of this instrument should point to this focus.

Figure 3 displays the frequency of occurrence of terms and the number of categories related to each definition. The frequency of occurrence of terms in the definitions shows that there is a balance among the definitions. It can be noticed that three of the four definitions with greater frequency were created by organisms responsible for setting standards. This explains the wide range of those definitions, once those organisms tend to produce more complete definitions. Another important aspect is the fact that both definitions with the smallest term frequencies are also the oldest. In other words, the most recent definitions tend to be more complete because they benefit from the maturity of the concept.

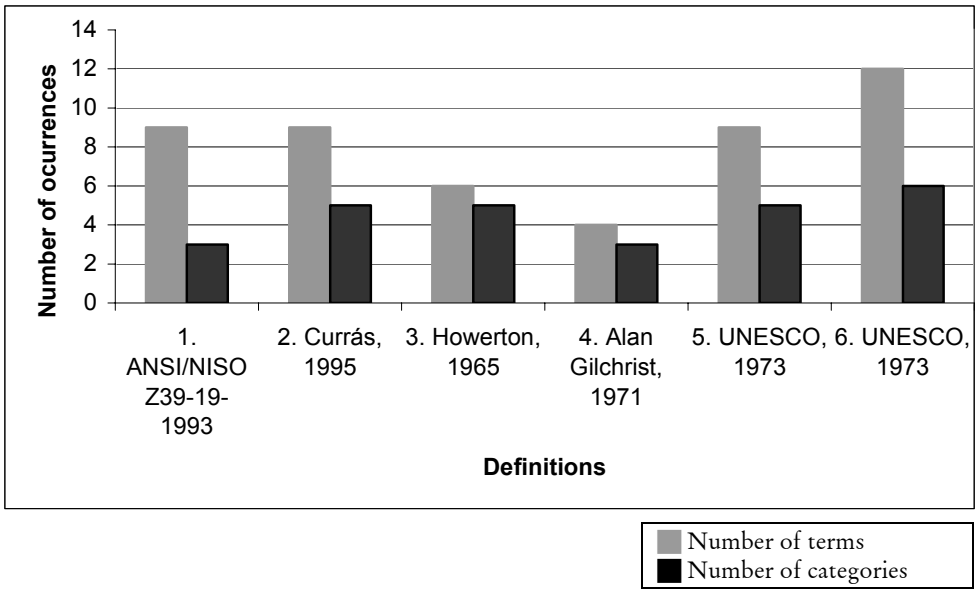

Figure 3. Occurrence of terms and categories by each thesaurus definition.

5.2 Comparative Analysis of the Occurrence of the Terms in the Definitions of Ontology and Thesaurus

In the thesaurus definitions can be observed a certain concern with the user (terms 'user' and 'users'), evidence of the relationship of the thesaurus with classification systems and document retrieval. The absence of occurrence of similar terms in the ontology definitions can indicate the difference of purposes between the two instruments. In the same way, the emphasis of the thesaurus definitions in the language category shows that the thesaurus focuses communication between users and classification systems. The terms "controlled," "normalized," and "authorized" evidence the terminological aspect of the thesaurus.

In the ontology definitions the terms "formal" and "Logic" reveal a certain concern with the rigor of the representation, what is comprehensible, once the ontology is used for information exchange between computational systems and, therefore, in order to rule out mistaken computations, it is necessary to have a clear specification of the descriptions. The thesaurus also presupposes the absence of ambiguities in the meaning of the terms; however, this condition is assured by the construction methodology.

The ontology emphasizes the formalization of the properties of the relationships due to the need of accomplishing inferences through a computer, while in the thesaurus the inference is accomplished by human interveners. The need for inference, in the case of ontology, has its roots in the automatic knowledge exchange between computer systems. In this situation, many times it is necessary to deduce, through inference rules, that a concept subsumes another or that they are under the same generic concept, so that matching of information can occur. In other situations ontology is used as a substratum for knowledge based systems equipped with inference mechanisms, where the need of formalization in a language with a rigorously defined semantics is a fundamental condition.

6. Conclusion

The thesaurus used in information science and the ontologies used in computer science possess different origins and purposes. The former was born as a practical instrument to aid indexing and searching of documents, and the latter from the need to describe objects and their relationships. It can be said that there are some contact points in those origins, once they are related to the description of some entity: the subject of a document in the first case, and objects and relationships in the second one. However, the differences also left their marks, influencing the final form of each instrument. In computer science the situation is a little fuzzier. It seems that everything that models a segment of reality can be called an ontology – it has become a buzzword. In this case, even the thesaurus can be framed as a terminological ontology.

The quantitative analysis showed the difference of purposes between the two instruments. The frequency of occurrence of terms, as well as the inclusion of the terms in the categories showed that the thesaurus has the purpose of serving as an instrument of terminological registration and being used by people, and not the aim of registration of knowledge for computer inferences. In the case of the ontology definitions, the occurrence of such terms as "formal" and "logical" indicates the need of domain knowledge registration in a language that can be processed by the computer for the accomplishment of inferences.

Further evidence of this conclusion is the fact that, using languages for ontology registration (e.g. OIL), it is easier to register certain properties of the relationships than using a thesaurus. However, this difference of expressiveness is not significant for the indexing task or search of documents.

Ontology as a system of categories, just as it is seen in philosophy, occurs in information science during the elaboration of a group of categories that will be used to organize information classification and recovery systems. On the other hand, in computer science this point of view is not adopted. Some researchers state this distinction explicitly (Valente 1995 and Guarino 1998). Guarino (1998, 2) states:

> In the philosophical sense, we may refer to an ontology as a particular system of categories accounting for a certain vision of the world. As such, this system does not depend on a particular *language*: Aristotle's ontology is always the same, independently of the language used to describe it. On the other hand, in its most prevalent use in AI, an ontology refers to an *engineering artifact*, constituted by a specific *vocabulary* used to describe a certain reality, plus a set of explicit assumptions regarding the *intended meaning* of the vocabulary words.

Ontology as viewed by computer science, is a system of concepts, as thesaurus is, and thus it belongs to the epistemological level and not to the ontological. The distinction between thesaurus and ontology occurs in the language used, in the level of formalization and in its purposes. In this sense, thesauri can be framed as ontologies. However, we suggest that this classification should not be adopted in information science, once the ontology, in the philosophical sense of a system of categories, has already been used in the scope of information science.

Some researchers agree with the similarity between the thesaurus and the ontology of computer science, but they allege that the distinction between them would be in the fact that ontologies allow a larger variety of relationships. Such vision should not be accepted and it is based on a misunderstanding of what a term is and what relationship is according to the thesaurus theory. The thesaurus, as well as some languages for ontology representation, presents a group of predefined relationships to be used for structure of concepts. This set of structuring relationships varies from thesaurus to thesaurus, depending on the aim and on the underlying theory. In this case, the relationships observed in the domain are represented in the thesaurus as well as any other concept, while in the computer science ontologies, the relationships are represented differently from the properties and restrictions, and structural properties (e.g. transitivity) can be attributed to them, which can be used in the accomplishment of inferences.

The conclusions of this work, supported by the analysis of the literature and by the quantitative analysis, are the following: 1) ontology of philosophy and ontology of computer science are different objects; 2) ontology of computer science and thesaurus are objects that operate on the same level, in other words, at the epistemological level; 3) ontology and thesaurus possess different purposes, and the first is directed to domain concepts registration aiming at automated inference while the second is directed to communication between the user and documentary languages; 4) the thesaurus accomplishes part of the objectives that computer science intends to with ontology and because of that they are named terminological ontologies.

References

Alvarenga, Lidia. 1993. Definição de publicações oficiais. *Revista da Escola de Biblioteconomia da UFMG (Universidade Federal de Minas Gerais)* 22: 213-238.

Ander-Egg, Ezequiel. 1978. *Introducción a las técnicas de investigación social: para trabajadores sociales.* 7.ed. Buenos Aires : Humanitas.

ANSI Z39.19:1993. 1993. *American national standard guidelines for thesaurus structure, construction, and use*. New York.

Campos, M.L.A. 2001. *Linguagem documentária: teorias que fundamentam sua elaboração*. Niterói: Eduff.

Chandrasekaran, B., Josephson, J.R. and Benjamins, V.R. 1999. What are ontologies, and why do we need them? *IEEE Intelligent Systems* 14(1). 20-26.

Cranfield, S., Purvis, M. and Nowostaski, M. 2000. Is it an ontology or an abstract syntax? Modeling objects, knowledge and agent messages. In W. Horn, ed., *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI 2000).* Amersterdam: IOS Press, pp. 16.1-16.4.

Curras, Emilia. 1995. *Tesauros: linguagens terminologicas,* trans. Antonio Felipe Correa da Costa. Brasilia: Instituto Braseleiro de Informação em Ciência e Tecnologia.

Dahlberg, Ingetraut. 1978. A referent-oriented, analytical concept theory for interconcept. *International classification* 5: 142-151.

Duarte, M.M. 2002. *Ferramenta de apoio à identificação de fatores críticos de sucesso em uma organização*. Belo Horizonte: Universidade Federal De Minas Gerais. Dissertação, Mestrado Em Ciência Da Computação.

Eriksson, H., Puerta, A.R. and Musen, M.A. 1994. Generation of knowledge-acquisition tools from domain ontologies. *International journal of human-computer studies* 41: 425-453.

Fensel, D., Horrocks, I., Van Harmelen, F., Decker, S., Erdmann, M. and Klein, M. 2000. OIL in a nutshell. In R. Dieng and Olivier Corby, eds., *Lecture notes in computer science volume 1937, Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management,* pp. 1-16.

Fensel, D., Van Harmelen, F., Horrocks, I., McGuinness, D.L. and Patel-Schneider, P.F. 2001. OIL: An ontology infrastructure for the semantic web. *IEEE Intelligent Systems* 16(2): 38-45

Foskett, D. 1985. *Subject and information analysis.* New York: Marcel Dekker, pp. 270-316.

García Morente, M. 1964. *Fundamentos da filosofia: lições preliminares*. São paulo: Mestre Jou.

Gilchrist, Alan. 1971. *The thesaurus in retrieval*. London : Aslib.

Gomes, H.E. 2004. *Classificação, tesauros e terminologia: fundamentos comuns.* www.conexaorio.com/biti.

Gomes, H.E. et al. 1990. *Manual de elaboração de thesaurus monolíngües*. Brasília: CNPQ/PNBU.

Gruber, T.R. 1992. *Ontolingua: a mechanism to support portable ontologies.* Stanford: Knowledge Systems Laboratory, Stanford University.

Gruninger, M. and Fox, M.S. 1995. Methodology for the design and evaluation of ontologies. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Montrèal, Quèbec, Canada, August 20-25, 1995.* San Mateo, CA: Distributed by Morgan Kaufman, pp. 6.1-6.10.

Guarino, N. and Giaretta, P. 1995. Ontologies and knowledge bases: towards a terminological clarification. In N. Mars, ed., *Towards very large knowledge bases: knowledge building & knowledge sharing*. Amsterdam: IOS Press, pp. 25-32.

Guarino, N. 1995. Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human-Computer Studies* 43: 625-640.

Guarino, N. 1998. Formal ontology and information systems. In *Formal ontology in information systems: Proceedings of the First International Conference (Fois'98), June 6-8, Trento, Italy*. Amsterdam: IOS Press, pp. 3-15.

Guarino, N., Masolo, C. and Vetere. G. 1999. ONTOSEEK: Content-based access to the web. *IEEE Intelligent Systems & Their Applications* 14(3). 70-80.

Hovy, E. 1998. Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In Antonio Rubio, ed., *First International Conference on Language Resources & Evaluation (LREC), Granada, Spain, 28-30 May 1998: Proceedings*. Paris: European Language Resources Association, pp. 535-542.

Howerton, Paul. 1965. Organic and functional concepts of authority files. In: Newman, Simon M. ed. *Information systems compatibility*. Washington: Spartan Books.

Jasper, R. and Uschold, M. 1999. A framework for understanding and classifying ontology applications. In Thomas L. Dean, ed., *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99), Stockholm, Sweden, July 31-August 6, 1999*. San Francisco: Morgan Kaufman, pp. 11.1-11.12.

Lancaster, F. W. 1986. *Vocabulary control for information retrieval,* 2. ed. Virginia: IRP.

Lancaster, F. W. 1987. *Construção e uso de thesaurus*: curso condensado. Brasília: Ibict.

Lawrence, Steve S., Bollacker, K., Lee Giles, C. 1999. Digital libraries and autonomous citation indexing. *IEEE computer* 32: 67-71.

Lenat, D.B., Guha, R.V. 1990. *Building large knowledge-based systems: representation and inference in the CYC project.* Massachusetts: Addison-Wesley.

Marconi, Marina de Andrade and Lakatos, Eva Maria. 1982. *Técnicas de pesquisa: planejamento e execução de pesquisas, amostragens e técnicas de pesquisa, elaboração, análise e interpretação de dados*. São Paulo: Atlas.

Mena, E., Kashyap, V., Sheth, A. and Illarramendi, A. 1996. OBSERVER: an approach for query processing in global information systems based on interoperation across pre-existing ontologies. In *Proceedings, First IFCIS International Conference on Cooperative Information Systems (CoopIS): Brussels, Belgium, June 19-21, 1996.* Los Alamitos, CA: IEEE Computer Society Press, pp. 14-25.

Mora, J.F. 1963. On the early history of ontology. *Philosophy and phenomenological research* 24: 36-47.

Poli, R. 2001. *Framing ontology.* http://www.formalontology.it/framing first.htm.

Russell, Stuart, Norvig, Peter. 1995. *Artificial intelligence: a modern approach*. New Jersey: Prentice Hall.

Swartout, B., Patil, R., Knight, K. and Russ, T. 1996. Toward distributed use of large-scale ontologies. In Brian R. and Mark A. Gaines, eds., *Proceedings of the 10th Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff, Canada*. <http://ksi.cpsc.ucalgary.ca/KAW/KAW.html>

UNESCO. 1973. *Guidelines for the establishment and development of monolingual thesauri*. Paris.

Uschold, M. 1996. Building ontologies: towards a unified methodology. In *Proceedings of the 16th Annual Conference of the British Computer Society Specialist Group on Expert Systems.* Edinburgh: Artificial Intelligence Applications Institute, University of Edinburgh, pp. 19.1-19.12.

Uschold, M., Gruninger, M. 1996. Ontologies: principles, methods and applications. *Knowledge engineering review* 11(2): 93-136.

Valente, A. 1995. *Legal knowledge engineering: a modeling approach*. Amsterdam: Ios Press.

Vickery, B.C. 1980. *Classificação e indexação nas ciências*. Rio de Janeiro: Bng/ Brasilart.

Villela, M.L.B., Oliveira, A.P., and Braga, J.L. 2004. Modelagem ontológica no apoio à modelagem conceitual. *Simpósio brasileiro de engenharia de software*. Brasília.

Welty, C., Guarino, N. 2001. Supporting ontological analysis of taxonomic relationships. *Data and knowledge engineering* 39: 51-74.

Wiederhold, G. 1994. Interoperation, mediation, and ontologies. In *Proceedings of FGCS '94 : Fifth Generation Computer Systems Workshop on Heterogeneous Cooperative Knowledge Bases.* Tokyo, Japan: Ohmsha, pp. 33-48.

"Thesaurus" definitions taken from the information science literature

"A thesaurus is a controlled vocabulary arranged in a known order and structured so that equivalence, homographic, hierarchical, and associative relationships among terms are displayed clearly and identified by standardized relationship indicators that are employed reciprocally." (ANSI/NISO Z39-19-1993)

"Thesaurus is a specialized, normalized, post-coordinate language used for documentaries means, where the linguistic elements that composes it – single or composed terms – are related among themselves syntactically and semantically." (Translated into English by the authors from the original in Portuguese: Currás 1995, 88.)

"[...] an authority file, which can lead the user from one concept to another via various heuristic or intuitive paths." (Howerton 1965 *apud* Gilchrist 1971, 5)

"[...] is a lexical authority list, without notation, which differs from an alphabetical subject heading list in that the lexical units, being smaller, are more amenable to post-coordinate indexing." (Gilchrist 1971, 2)

[....] "a dynamic controlled vocabulary of terms related semantically and by generic relation covering a specific knowledge domain." (Translated into English by the authors from the original in Portuguese: UNESCO 1973, 6.)

[....] "a terminological control device used in the translation of the natural language of the documents, from the indexers or from the users in a more restricted system language (documentation language, information language)." (Translated into English by the authors from the original in Portuguese: UNESCO 1973, 6.)

"Ontologies" definitions taken from the computer science literature

"[....] ontology is a representation vocabulary, often specialized to some domain or subject matter." (Chandrasekaran et al. 1999, 1)

"[...] ontology is sometimes used to refer to a body of knowledge describing some domain, typically a commonsense knowledge domain, using a representation vocabulary." (Chandrasekaran et al. 1999, 1)

"An ontology is a declarative model of the terms and relationships in a domain." (Eriksson et al. 1994, 1)

"[...] an ontology is the (unspecified) conceptual system which we may assume to underlie a particular knowledge base." (Guarino and Giaretta 1995, 1)

Ontology as a representation of a conceptual system via a logical theory". (Guarino and Giaretta 1995, 1)

"An ontology is an explicit specification of a conceptualization." (Gruber 1993, 1)

"[...] An ontology is a formal description of entities and their properties, relationships, constraints, behaviors." (Gruninger and Fox 1995, 1)

"An ontology is set of terms, associated with definitions in natural language and, if possible, using formal relations and constraints, about some domain of interest …" (Hovy 1998, 2)

"Each Ontology is a set of terms of interest in a particular information domain, expressed using DL …" (Mena et al. 1996, 3)

"[...] An ontology is a hierarchically structured set of terms for describing a domain that can be used as a skeletal foundation for a knowledge base." (Swartout et al. 1996, 1)

"An ontology may take a variety of forms, but necessarily it will include a vocabulary of terms and some specification of their meaning." (Uschold 1996, 3)

"Ontologies are agreements about shared conceptualizations." (Uschold and Grunninger 1996, 6)

"[....] a vocabulary of terms and a specification of their relationships." (Wiederhold 1994, 6)