Using Topic Maps in Establishing Compatibility of Semantically Structured Hypertext Contents

Guilherme Baião Salgado Silva* and
Gercina Ângela Borém de Oliveira Lima**

* Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627 Sala: 4053 – Pampulha
31270-010 Belo Horizonte/MG, Brazil <guilherme.baiao@gmail.com>
** Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627 Sala: 4053 – Pampulha
31270-010 Belo Horizonte/MG, Brazil <glima@eci.ufmg.br>

Guilherme Baião Salgado Silva holds a Master's in Information Science (UFMG, Brazil) and a Bachelor's in computing science (*Pontifícia Universidade Católica*, Brazil), and is a specialist in database (*Centro Universitário de Belo Horizonte*, Brazil). He works as a systems analyst-programmer and manager of information databases of the Court of Justice of Minas Gerais State (Brazil).

Gercina Ângela Borém Lima is full professor of information science at the Information Science School – UFMG (Brazil). She holds a PhD in Information Science from UFMG (Brazil) and a Master's in Library Science from Clark Atlanta University (USA). Currently, she is the vice president of ISKO's Brazilian Chapter, the coordinator of the Graduate Program in Information Science at UFMG. She has been funded by a grant from the research agency CNPq to coordinate the Research Group MHTX (models for hypertext organization) since 2010. Current research includes the areas of knowledge organization and hypertext systems, both applied to digital library.

Baião Salgado Silva, Guilherme, and Lima, Gercina Ângela Borém de Oliveira. **Using Topic Maps in Establishing Compatibility of Semantically Structured Hypertext Contents.** *Knowledge Organization. 39(6)*, 432-445. 30 references.

ABSTRACT: Considering the characteristics of hypertext systems and problems such as cognitive overload and the disorientation of users, this project studies subject hypertext documents that have undergone conceptual structuring using facets for content representation and improvement of information retrieval during navigation. The main objective was to assess the possibility of the application of topic map technology for automating the compatibilization process of these structures. For this purpose, two dissertations from the UFMG Information Science Post-Graduation Program were adopted as samples. Both dissertations had been duly analyzed and structured on the MHTX (Hypertextual Map) prototype database. The faceted structures of both dissertations, which had been represented in conceptual maps, were then converted into topic maps. It was then possible to use the merge property of the topic maps to promote the semantic interrelationship between the maps and, consequently, between the hypertextual information resources proper. The merge results were then analyzed in the light of theories dealing with the compatibilization of languages developed within the realm of information technology and librarianship from the 1960s on. The main goals accomplished were: (a) the detailed conceptualization of the merge process of the topic maps, considering the possible compatibilization levels and the applicability of this technology in the integration of faceted structures; and (b) the production of a detailed sequence of steps that may be used in the implementation of topic maps based on faceted structures.

Received 10 July 2010; Revised 14 August 2012; Accepted 23 August 2012

1.0 Introduction

The growing use of hypertext systems through the Web has triggered a series of changes in information organization, storage, and retrieval concepts. With their non-linear associative structure, hypertext documents offer a number of innovations in relation to traditional printed documents, mainly in regard to research and learning processes, since hypertextual reading is predominantly based on navigation. On the other hand, hypertext systems have some disadvantages for users, including cognitive overload and disorientation during navigation.

With the goal of mitigating such problems, several hypertext structured navigation systems have been proposed. These new models are based on the application of techniques and methodologies for the organization of information in the hypertextual semantic network and on interfaces that allow for oriented navigation supported by the ability of visualizing the global structure of the network in a number of ways, such as hierarchical diagrams and conceptual maps. Some projects, like the Hypertextual Map (MHTX) proposed by Lima (2004), suggest the application of the theory of facet analysis in the analysis of traditional documents and their conversion into hypertextual documents for later availability in a digital library of theses and dissertations. However, the hypertextual documents in these repositories have heterogeneous semantic structures, unified search requires the identification of interrelationships between document structures, a rather complex process.

Therefore, the problematic issue to be dealt with in this study is characterized by the existence of distinct conceptual structures for the representation of different hypertextual documents in one same repository (a digital library, for example), and which may approach subjects that are correlated among themselves or even approach the same subject through different terms and contexts. For that reason, the creation of mechanisms that enable the interrelationship between these structures tends to facilitate the retrieval of information by users.

Considering that the subject matter analysis of each document was conducted based on theory of facet analysis and supported by literature warranty in relation to the documents analyzed, at the end of the process we shall have different controlled vocabularies being used for the organization of documents. These differences occur mainly due to the possibility of designating different terms for the same subject and to the possibility of one subject being associated with others, in distinct ways, in each of these structures.

This study is intended to assess, in the light of theories for the compatibility of languages and vocabularies, topic map standard as a tool for the automated integration of heterogeneous faceted conceptual structures in hypertextual contents. These maps may be used for the representation of information using Knowledge Organisation Systems (KOS), such as thesauri, taxonomies, conceptual maps, and faceted classifications (Garshol 2004).

2.0 Topic maps

Topic maps form a standard that allows for the implementation of information representation instruments, a standard that is used to describe and navigate through informational objects in digital systems in a contextualized manner. According to Park (2003), topic maps and the organization of knowledge are natural partners, since the first one was invented so as to meet the requirements and solve problems with the latter, mainly in respect to the semantic integration (or compatibilization of heterogeneous subject schemes) and shared construction of conceptualizations (or mental models). For Park, the construction of topic maps must always be based on theories, techniques, and methodologies developed by knowledge organization for the construction of intellectual structures, such as faceted classification.

As in other information representation instruments (such as thesauri and concept maps, for example), topic maps are based on the representation of subjects (topics) through relations (associations). Based on the mapping of subjects and associations, "paths" that point to information resources that are relevant for these subjects are constructed. This paradigm requires that the topic map's author think in terms of topics (subject matter, conversation topics, specific notions, ideas or concepts) and associate various types of information to a specific topic. For example, a topic called "hardware" can be associated with another "topic" called "printer" throught an association of type "is a." The topic "printer," in turn, can be linked to different kinds of information resources (occurrences), like an image of a printer or a text that describe its functionality.

Lima and Fagundes (2004) note that the original application of this language was the construction of indexes and glossaries for documents, which was later extended to the Web for the definition of relations between entities, constituting one structure that makes

data search more efficient. Librelotto (2005) lists the following as main objectives of topic maps:

a) to structure information resources by means of mechanisms that are external to these resources;
b) to allow for searches that retrieve the desired information; and,
c) create different views for users or specific applications by means of information filters.

Garshol (2002) observes that, through topic maps, it is possible to create external indexes to describe information that is present in documents or databases.

According to Ahmed and Moore (2006), the separation between the conceptual structure and indexed resources presents three remarkable benefits:

a) the map serves as one representation, at the highest level, of the indexed information, thus facilitating the location of resources and the conceptual structuring of those resources being represented;
b) ease of topic map fragmentation so as to meet different requirements; and,
c) ease of combination among topic maps that index different sets of resources.

The topic map standard was initially conceived in the beginning of the 1990s. Since then, several research studies have been conducted with the goal of creating norms and standards for their use, as well as of assessing their applicability. In the late '90s, the ISO/IEC group created an international standard ISO 13250 for the formalization of knowledge structures represented by this technology. Soon after the publication of ISO 13250, studies were begun aimed at its adaptation to XML (eXtensible Markup Language), which already stood out in terms of acceptance as a language for communication among information systems on the Web (Park and Hunting 2003). At that same time, the independent organization *TopicMaps.org* was founded with the goal of creating an XML specification for topic maps. XTM 1.0 (*XML Topic Maps*) appeared in the beginning of 2001.

According to Pepper (2000), the topic map standard was created based on three main components (also present in the bibliographical indexes): topics, associations, and occurrences. In addition to these three, Pepper also highlights other important elements in the topic maps: subject identity, facets, and scope. Due to the mnemonic composition of these components, the first group was named TAO by Pepper and the second group, IFS. Comments will be made on each of these concepts below.

A topic is the representation of a subject. More generically, a subject could be anything: a person, a concept, an entity, etc. In an ideal representation, the relation between topics and subjects is of a "one-to-one" type; that is, each topic represents one and only one subject, and each subject is represented by a single topic.

Within the topic map, any topic is an instance of one or more "topic types." For instance, it can be said that the "indexer" topic is of type "Person" (the type "Person" being one topic in itself), and that the "subject matter analysis" and "translation" topics are of the "Process" type and "book" is a "material" type topic.

For each topic in the map, it is possible to define one or more topic names. Librelotto (2005) notes that the possibility of attributing various names to one topic may have two objectives: to facilitate the understanding of the meaning of the topic through alternative descriptions and to indicate the use of different names in different contexts, like language, domain, geographical area, historical period, and so on. According to Park (2003), the utilization of multiple names for one same topic is a fundamental requisite for the robustness, scalability, and interoperability of maps.

In a topic map, associations are meant to express a relationship between one or more topics. An association of topics is, formally, a linking element that describes the relations between two or more topics (Pepper 2000) and the roles each of them play in that association. Associations between topics may also be grouped according to their association types, which, in turn, are also defined through topics. In these associations as well, as cited in Pepper (2000), we can define also the role of each topic. For this purpose, the association roles element is utilized. For example, when there is a need to represent, in a topic map, a whole-part type relation between subject matter analysis processes and subject extraction, it is possible to define not only the type of relation (whole-part), but also to define that the whole is the subject matter analysis and the part is the extraction of subjects.

Contrary to what happens in thesauri and in other traditional classification schemes, topic maps allow for the representation of any type of relation between subjects. The class-instance type relation is established by typifying topics, associations, or occurrences (Pepper 2000). For the remaining types of relations, new associations must be defined.

In topic maps, an occurrence is any information that, in some manner, is specified as being relevant for a certain topic (subject). That is, occurrences allow a topic to be related to any number of resources considered to be relevant for it. These resources are then called topic occurrences. Ruiz (2005, 86) defines occurrence as "external information resources, linked by a reference that is useful for its location, which define or exemplify the meaning of the topic." For Pepper (2001), occurrences may be external to the topic (one Web page about that topic, for instance) or internal (defined in the map proper). As with topics and associations, occurrences may also be classified by means of occurrence types. Therefore, it is possible, for example, to distinguish those occurrences that define the topic from those occurrences that exemplify or instanciate the topic.

As mentioned before, in the second group of topic map components defined by Pepper (2000), we will find the element responsible for the identification of subjects, facets, and scope.

The very first of them, the subject identifier, is intended to guarantee that a subject is represented by one single topic in the map. For that purpose, the subject can be identified in the topic map through:

a) already existing and addressable electronic resources, such as an image on a webpage, for instance, or
b) a subject indicator created and made available for this specific purpose, since the subject in question is not addressable, such as a subject "Brazil," for instance.

In the latter case, the resource created by a community with the specific goal of providing a subject with one single identification is called a published subject indicator or simply PSI. The Topicmaps.org site (http://www.topicmaps.org/xtm/), for example, defines a series of PSIs related to electronically non-addressable subjects, such as languages, countries, and the subjects themselves that are involved in the construction of topic maps. The second component, the facets, is in fact metadata used to describe occurrences within a topic map. This resource was adopted by standard ISO 13250 and eliminated by the XTM standard, since, in the latter, the resource and the property value must be considered topics that are associated by means of "applicable to" type associations (Park 2003).

The scope is used in topic maps to define the context in which topic characteristics are valid, removing ambiguities and reducing the chance of errors when merging topic maps (Pepper 2001). Topic characteristics are understood to be the names, associations, and occurrences in which the topic takes part. According to Ahmed and Moore (2006) "scope is the term used in the topic map norm to refer to a restriction or a context in which something is said about a topic." If no scope is attributed to the name, to the occurrence, or to the association, this means that all topic characteristics are valid in any situation. The merge process is a resource made available by topic maps to integrate two maps into one, in a manner that guarantees that topics that represent the same subject in the two maps are transformed into one single topic in the resulting map. When this merging takes place, the characteristics of the merged topics are to be unified, thus eliminating duplication, and kept in the resulting topic (names, occurrences, associations, and identifiers).

This process can be automated through applications that are specific for the manipulation of topic maps. These applications use two distinct possibilities to determine that two topics represent the same subject. The first one occurs when the single subject identifier is the same for two distinct topics, whether the identifier is an addressable object or is a PSI. The second occurs when there are one or more names of a topic that coincide with one or more names of another.

The second case requires that the homonymy problem be taken into account, that is, when one same term represents two or more different subjects. In order to deal with this problem, applications can take two different paths: in the first situation, the application presents to the user those topics that were identified with the same name so that the user may decide whether the topic should be merged or not. In this case, the process is no longer totally automated; in the second situation, the application takes into account the equivalence of the names only if they are defined for the same scope, thereby turning it into a fully automated process.

Another merging possibility among map elements is the use of the mergemap component to specify the topics that are to be considered as being one only. The mergemap refers to an external map through a URI, thus allowing for the indication, in the map topics, of corresponding topics in other maps.

Since some topic maps may contain a great quantity of topics and associations, it is currently possible to find tools that enable the visualization of topic maps through different interfaces: textual, generally made available through HTML pages, graphs, trees or maps

of several different types, including tridimensional types. *Omnigator*, for instance, enables the visualization of maps both in page format and in graphic format (http://www.ontopia.net/omnigator/). Another example is the *TMNav*, which allows for the visualization of maps in several different graphic forms, among which is the tree form. *TMNav* is one of the tools available in the *TM4J* (http://tm4j.org/), which is an application package for the creation, management, and visualization of topic maps. The most recent studies about topic maps concentrate on the implementation of resources aimed at making the definition of restrictions in relationships[1] and inference rules viable.

3.0 Compatibilization of languages and controlled vocabularies

Studies concerned with the compatibilization of languages were started in the 1960s (Dahlberg 1981). Dahlberg (1981) highlights the work conducted by William Hammond and Staffan Rosenborg, Simon M. Newman, Madeline M. Henderson, John S. Moats, and Mary Elizabeth Stevens. One of the main results reached by the studies at that time was the definition of some core concepts for the field, such as convertibility, compatibility, and their respective applications, mainly within the scope of cooperation among information centers. William Hammond, in 1965, defined compatibility as: "the capacity of an information system to accept indexing data and summarization from another system on any subject matter that is common to both" (Newman 1965, 7).

According to Maniez (1997) the words "integration," "harmonization," "reconciliation," and "concordance" are usually applied to indicate the concept of convergence, that is, the same concept of "compatibilization." However, Lancaster (1986) considers integration of vocabularies as a specific method of compatibilization, in which there is no conceptual analysis, that is, for a certain user request, all related occurrences (including word variations) and elements are shown as a result, without showing the relation or the context in which each one of them are present.

Coates (1970) comments that up until then cooperation among libraries was based solely on the union of catalogues, that is, there were no efficient mechanisms that allowed for the exchange of descriptions of a subject matter. However, according to Maniez (1997), the 1970s witnessed an increase in the production of studies on language compatibility. Some relevant authors of that time cited by Dahlberg (1978) were, among others: Hans Wellisch, Verina Horsnell, Gernot Wersig, Eric Coates, Elaine Svenonius, Dagobert Soergel, Linda C. Smith and V. M. Glushkov. In addition to these authors, Lancaster (1986) points out the work of H. Neville and R. T. Niehoff, cited in Soergel (1974).

In 1971, UNESCO's UNISIST report dedicated an entire section to the conversion and compatibility of indexing instruments. The report defines compatibilization and conversion as follows (Dahlberg 1981, 86):

> The report defines 'compatibility' as the quality of systems whose products may be used in an interchangeable manner, despite the differences in notation, structure, physical support, etc., with no special automated conversion; and 'conversion' as the process through which information entries are transformed based on code transcription, data structure, etc., thus allowing for them to be interchangeable between two or more services or systems that utilize different conventions and media.

In the following decade, the 1980s, there was little research into the theme (Maniez 1997, 213). At that time, technologies that allowed for the interchange of data among different bases were rather restricted. In the 1990s, studies were then resumed due to the global expansion of information networks, the ease of simultaneous access to different collections and the increased amount of interchange among different databases, most importantly over the Web (Maniez 1997; Hudon 2004). These studies presented new definitions for compatibility. Noteworthy among them is the definition provided by Gerhard Riesthuis in a seminar on system integration and compatibilization held in 1995, in Varsovia: compatibility means that, for each term in a language there is a corresponding term with the same meaning in another language, and that is why it is possible to convert terms of one language into another, with no alterations in meaning (Maniez 1997). Maniez (1997) also comments that the compatibilization of languages is a crucial aspect in any information retrieval system, since, for these systems, users are required to use a language that is compatible with that which was used for indexing the entries, considering the subjectivity implicit in the indexing process.

In Brazilian literature, among studies produced since the beginning of the current decade, we highlight those conducted by Dr. Maria Luiza de Almeida Campos. Campos (2005) cites as the main language compatibility theorists in information science: Soer-

gel (1982), Dahlberg (1981), Neville (1970), and Glushkov et al. (1978).

According to Zhang (2006), the objective of promoting the compatibility of indexing languages is to enable users, through a single search strategy, to retrieve information stored in any of the bases or repositories that comprise a certain information retrieval system. For Rada (1987), the integration[2] of controlled indexing languages may come to serve for different purposes:

a) to enable users to use a term for a certain subject and, based on that, allow for the indication of appropriate descriptors in each base;
b) to permit the indexing process to be shared through the possibility of converting one thesaurus into another;
c) to make the identification of relations between different controlled vocabulary subjects possible; and
d) to increase the vocabulary content.

In 1965, William Hammond had listed the following compatibilization objectives (Newman 1965):

a) to provide for the dissemination of information resulting from ongoing searches;
b) to eliminate the need for doubled indexing and for search report summaries; and
c) to furnish for the retrieval of reports stored in different repositories, based on the original indexing.

For Guinchat and Menou (1994), the compatibility between documentary languages is of utmost importance for the exchange of information between units that deal with the same or interrelated subject matters. According to Soergel (1982), the main obstacles found in the controlled vocabulary compatibilization process are:

a) different levels of precombination of the descriptors used by each of the systems;
b) different degrees of specificity for subjects that compose each one of the systems;
c) different ways to relate one specific term to one single generic term in the case of polyhierarchies; [3]
d) different possibilities for the aggregation of specific subjects formed by terms that, in an isolated manner, do not represent subjects;
e) the possibility of different connotations for one same term (homonyms); and
f) the lack of subjects in some of the systems, although this same subject occurs in other systems involved. This last obstacle tends to be bigger as the overlaying of subject matters dealt with by those vocabularies involved decreases.

In addition to these problems, Lancaster (1986) comments on problems related to the use of computational processing in an attempt to automate the compatibilization process and notes, as the main issue, the possibility that one same term may have different meanings. For the same author, another problem is the existence of errors in the form in which the terms are written in some systems. For Doerr (2001), the difficulties found in the compatibilization process originate mainly in the subjectivity of the choice of descriptors and hierarchical relations during the construction of controlled vocabularies.

The compatibilization of controlled vocabularies can occur at three levels: terminological (or verbal), conceptual, and structural (Soergel 1982). At the first one, the terminological level, procedures are based on the comparison of names or descriptors that have been attributed to the subjects. The second level, conceptual, may be considered the most complex of all, since it is at this level that nearly all problems previously mentioned are found. Due to this complexity, contrary to what occurs at the terminological level, compatibilization hardly ever occurs in an automated manner at the conceptual level. Also at the conceptual level, methods for the validation of the previous stage may be proposed, in view of the fact that homonyms may generate erroneous terminological compatibilization from the conceptual point of view. In turn, compatibilization at the structural level is responsible for guaranteeing the conformity of those relations established among elements.

Hudon (2004) mentions a fourth level, the compatibilization of subject matter. This level is related to the possibility of representing the same subject matters in different controlled languages, whether by means of a single element or through the combination of two or more of them. Glushkov et al. (1978), in turn, divided the compatibilization of vocabularies into two levels: the semantic level, linked to the capacity of both to represent the same body of knowledge, and the structural level, related to the similarity according to internal characteristics, which corresponds to the terminological, conceptual, and structural levels mentioned by Soergel (1982).

In the compatibilization model proposed by Dahlberg (1981), for instance, the three noticeable stages

of compatibilization are: verbal (alphabetical comparison matrix constructed in topic 4 of the author's study), conceptual (topics 5.1 to 5.4) and structural (5.5 to 5.7).

We found different techniques and algorithms for the integration of languages, among which we may highlight mapping, the use of intermediary languages, microthesauri, macrothesauri, and the universal thesaurus. Among these, mapping is the one most often found in references to the subject.

The mapping consists of determining, for each term of a vocabulary, the term with the closest correspondence in another vocabulary. This mapping may be unidirectional or bidirectional. Unidirectional mapping identifies, for those terms in a first vocabulary X, those with the closest possible equivalence in a second vocabulary Y. Therefore, all terms from X will have a correspondent in Y, but not the other way around. For the reverse order also to take place, it is necessary to undertake bidirectional mapping, that is, also map those equivalent terms in X for each term in Y. According to Doerr (2001), mapping is the basic process in all other compatibilization techniques.

As for the number of vocabularies involved increases, the situation becomes extremely complex, and a more viable alternative is then the use of intermediary languages. In this case, each vocabulary must be mapped bidirectionally with the intermediary language and, in so doing, produce the correspondence among all vocabularies involved.

With respect to the main difficulties found during mapping (whether intermediary languages were used or not), Lancaster (1986) points out:

> a) differences between pre-coordination levels of involved vocabularies;
> b) differences between levels of specificity; and
> c) the lack of corresponding elements in one of the vocabularies.

For both the first two situations, the solution proposed by the author is to map one element of a vocabulary in relation to various elements in the other. Eventually, the solution would be to add the element into one of the vocabularies. This situation occurs in smaller proportions as the overlaying of subject matters covered by both vocabularies increases.

Neville (1970) presents an example of a step-by-step mapping of thesauri, which he denominated reconciliation of thesauri. Considering the terminological and conceptual models, the author identified six possible correspondence types between elements from distinct vocabularies: (1) exact correspondence, (2) synonyms, (3) generic to specific, (4), mappable to a different pre-coordination level, (6) semantic factoring, and (6) antonyms. It is important to note, however, that Neville (1970) did not consider the treatment of homonyms, except in those cases in which at least one of the vocabularies involved construes the occurrence of homonymy, as well as, for example, a case in which a vocabulary distinguishes the elements "tank (war vehicle)" and "tank (container)" and the other vocabulary presents only the term "tank." Also with respect to homonymy, it is worth noting that it is a type of correspondence that is not as common when we are working with vocabularies from the same knowledge field, as was done in the study by Neville.

For Tennis (2001), the work of Neville (1970) represents a specific type of mapping, the supra-thesauri, which may be defined as a type of mapping that utilizes intermediary language, in which this language is generated after an iterative process of conceptual analysis of the involved thesauri. Therefore, the supra-thesauri only exist after the analysis of a predefined set of thesauri, and that is why its application is limited to this group. This procedure is not supported by any literary warrant, as it is replaced by the conceptual warrant derived from the involved vocabulary.

Soergel (1974) presents a table for the mapping of controlled vocabularies in which he suggests five possible levels of conceptual equivalence among elements of distinct languages and, for each of these levels, five possible levels of terminological equivalence: same term, distinct terms, and descriptor combinations (utilizing logical operators). The table proposed by Soergel (1974) is presented in Table 1 below.

The mapping of one vocabulary into another is a tiresome and time-consuming intellectual task (Lancaster, 1986). For this reason, some studies were conducted in the late 1960s and early 1970s with the purpose of developing algorithms that would automate at least a part of the process. Some types of mappings that may be realized automatically were identified in these algorithms.

Lancaster (1986) also produces another specific type of technique for the compatibilization of vocabularies which he calls the integrated vocabulary approach and whose purpose is slightly different from that of intermediary languages, as equivalences among elements of involved vocabularies are defined according to the user's search strategy and, consequently, no new neutral language is created. In those

Expression of searching equivalent: to the A-descriptor corresponds		Degree of equivalence				
		Precise searching equivalence in B	Approximate Searching Equivalent in B			No searching equivalent in B
			Broader searching equivalent	Narrower searching equivalent	Related searching equivalent	
One B-descriptor	Same term	A: Education B: Education	A: Food Service (in school) B: Food service (in general)		A: Educational assessment (of individuals) B: Educational assessment (of programs)	B: Movement (social, political) B: Movement (physical ed.)
	Different terms	A: Examination B: Test	A: Public University B: Higher Educational Institution	A: Selective Process B: University Selection Exam	A: Education efficiency B: Quality of schools	
A combination of B-descriptors	OR-combination of B-descriptors	A: Social Sciences B: Social Sciences OR Psychology OR Sociology	A: Church-sponsored educational institution B: Private Elementary School OR Private Secondary School OR Private Coll. & Univ.	A: Educational Opportunities B: Equal education OR Educational discrimination		
	AND-combination of B-descriptors	A: Entrance examinations B: Tests AND Admission	A: Oral entrance examination B: Tests AND Admission			
	Combination of B-descriptors using both AND and OR	A: Student relationship to authority B: Students AND Behavior AND (Educational instructors OR Parents)	A: Student rebellion against authority B: Students AND Behavior AND (Educational instructors OR Parents)			

Table 1. Terminological and conceptual compatibilization matrix Source: Adapted from SOERGEL, 1974, p. 506.

cases, the users themselves are the ones who select the comparison strategy: exact equivalence, synonyms, related terms, number of terms to be exhibited as results, etc. The search result displayed the occurrences at the selected data bases, each one with their specific terms and the reason for which the occurrence was considered. For Lancaster (1986), the compatibility matrix proposed by Dahlberg (1981) is another example of the integrated vocabulary approach.

4.0 Procedures adopted

Two faceted conceptual structures were utilized as the empirical object for the experiment. The first one was developed by Lima (2004) during the implementation of the MHTX prototype, presented in her doctorate dissertation. It is the result of the facet analysis of the doctoral dissertation by Madalena Martins Lopes Naves, PhD in Information Science from the UFMG Information Science School, from the year 2000. In that analysis, the set of principles suggested by the Spiteri Simplified Model was utilized.

The second one was constructed by Professor Gercina Ângela Borém de Oliveira Lima, in partnership with Professor Maria Luiza de Almeida Campos, both with degrees in librarianship and PhDs in information science. It resulted from the facet analysis of the doctoral dissertation by Lima (2004). In that case, the categorization of subjects and facets was based on formal categories and subcategories proposed by Dahlberg (1978).

It is worth noting that, in addition to the differences between the categorization models adopted in the construction of each one of the two schemes, the latter was constructed independently of the scenario used in the first one, that is, at no time was there any preoccupation with the compatibility between the two. Therefore, they can be considered an ideal sample for construing the problematic scenario previously described.

According to Garshol (2004), the three steps for the creation of a topic map based on a faceted classification system are:

a) create, for each facet, a "facet" type topic named after the facet;
b) create, for each highest level term of each facet, a "term" type topic, whose name will be the term proper, and associate it to the corresponding facet using a "belonging to facet" type association;
c) create, for each term just below the highest level term, a "term" type topic whose name will be the term proper, and associate it to its father using a "generic/specific" type association.

In order to maintain the representativeness of the conceptual maps developed by Lima (2004) and avoid semantic loss in the faceted structure, it was necessary to make some adaptations to the route suggested by Garshol (2004):

a) instead of creating one single "facet" type topic, eight topics were created, each of them representing one different facet or subfacet level, since, in the scheme proposed by Lima (2004), at each level a different symbol attribute was given in the conceptual map;
b) for each association roles were also defined for the topics involved. In a "belonging to facet" type association, for example, one of the topics involved takes up the "facet" role and the other one the "term" role;
c) in those cases in which synonymous terms were kept (Concept/Idea/Thought, for instance), the "topic names" property was used to indicate that all the terms referred to one same subject;
d) to indicate the division criteria, "division criteria" type topics were created and, subsequently, associated to the corresponding facet through a "divides the facet" type association.

In order to make a "belongs to the facet" type association between those terms subordinate to each of these criteria and the immediately superior facet, it was necessary to add a third role indicating the division criteria with the goal of making explicit which terms that are subordinate to the facet belong to which criteria. This means that, in the case of subordinations for which there was an explicit division criterion in the original scheme, a ternary relation was generated involving the facet, the subfacet (whose role was defined as "term") and the division criteria in question.

Following that, occurrences were created for each one of the topics. Each occurrence pointed to the HTML page corresponding to the part of Naves's (2000) dissertation that best defined its semantic content. Therefore, all links between elements of the map and the respective information resources that had been created by Lima (2004) were kept. From that point on, it was possible to navigate in context through the topic map.

The next step was the creation of a PSI repository. As previously mentioned, PSIs are especially responsible for the portability of topic maps, for they are intended to provide one singular identification of a subject with no ambiguities. For that reason, even though a PSI may be represented by any type of data, the ideal situation is that they be in a repository from which they are made available on the Web and that each of their elements be addressable by means of a specific URI. This allows for the reutilization and the collaborative construction of ever more complete and more comprehensive PSI repositories from the conceptual point of view.

In order to assemble a second topic map, the same procedures were adopted, this time based on a distinct faceted structure that was obtained from the analysis of the dissertation by Lima (2004). However, in this second map, the PSIs were not added, since the creation of a PSI for each new topic, as was done in the first map, might generate more than one PSI for the same subject, in the event a subject was being dealt with in both maps. That is, the reutilization of already published subject matters demands previous manual conceptual analysis to be performed by the professional constructing the topic map. Since this study deals with the compatibilization by means of automated resources already present in the topic maps, conducting the analysis for the insertion of PSIs in the second map would be justifiable.

The software used to aid in the construction, editing and navigation of the topic maps was Topincs <http://www.cerny-online.com/topincs/>, version 1.3.3, which is a free, open code software application for academic purposes and which allows for the exportation of maps into XTM files. This resource provides for the visualization of maps created in other topic map editing tools and navigators that support the same standard.

TMNav was used for graphic visualization of the maps. This software enables different graphic forms of visualization, such as hierarchical (in tree format) and through a hyperbolic map, which, according to Lima (2004) and Silva (2007) is a highly recommended interface for navigation in hypertextual resources.

Several other software programs were analyzed before these were selected. However, some presented a high level of installation and handling complexity, requiring, inclusively, the mastering of a programming language or code architecture (TM4J, available http://www.techquila.com/tm4j.html), requires broad knowledge of Java for the installation and compilation of its components). Others, like TMProc[4] do not support the XTM standard and require the installation of several other little-known software programs.

For the automated merging of the two topic maps produced during the previously described stages, the *Ontopia OKS Samplers* was used.

5.0 Results analysis

The first step of the results analysis was the manual unification of the faceted structures involved. The unified structure resulting from this procedure was used as a parameter for comparison with the topic map obtained after the automated merging of the two previously generated maps originating from these same structures. For the production of the manually unified scheme, the structure related to the dissertation by Lima (2004) was used as the basis. Following that, an attempt was made to allocate into this basic structure each of the descriptors from the second scheme, related to the dissertation by Naves (2000), considering the subject to which each one referred to in each piece of work separately.

During this process, it was possible to identify different types of relations between the elements of these two structures. These relations were classified considering the three levels of compatibilization (terminological, conceptual, and structural) presented by Soergel (1982). The first type of relation identified was the terminological, conceptual and structural equivalence, that is, the same descriptor present in both structures, thus denoting the same concept and in similar hierarchies.

A second type of relation was identified among the elements of the same terminology and conceptual load with slight variations such as number variation. Another type of relation was identified among those elements that were equivalent from the conceptual and terminological point of view, but were presented under a different subordination in each scheme, which implies a divergence at the structural level. The fourth type is a variation of the latter, due to terminological variations between descriptors. The fifth type was identified among elements with different descriptors and which were conceptually divergent only in relation to their specificity. The last type of relation identified occurred in those cases in which the elements had no previous relations, but had their allocation defined so as to provide for the structural compatibility between the schemes involved and the resulting scheme.

In order to provide for an easier identification of the types of relations in the resulting scheme, the following symbol mapping strategy was adopted:

Symbol	Description
Δ	Structural, conceptual and terminological equivalence
∓	Terminological variation; structural and conceptual equivalence
▲	Terminological and conceptual equivalence; structural divergence;
■	Terminological variation; conceptual equivalence; structural divergence;
◊	Terminological divergence; conceptual divergence only in relation to the level of specificity; structural equivalence;
□	Structural equivalence only;

Table 2. Symbols attributed to each of the relation types identified during manual unification.

Figure 1 shows a small part of the structure resulting from the unification as a function of the steps described previously.

The identification of the relations, besides allowing for a better comparison of results, serves as a justification for the insertion of each term from the second scheme into the first scheme. An example of that is the term "exact sciences" that, as it had been inserted in the symbol ◊, allows us to deduce that there is a divergence only in the specificity level. Considering that the hierarchical chain in which the term now takes part, one may deduce that it is a subject whose specificity level is greater than that of "Natural Real Sciences" and smaller than "Computer Sciences."

In order to facilitate the comparison of results from both processes (the manual unification of the two schemes and the automated merging of the topic maps), it was necessary to navigate throughout the resulting topic map using *OKS Samplers* so as to extract a hierarchical structure in the same format as that of the scheme generated during the manual unification.

At this point, it was noticed that there were topics in the resulting topic map that were subordinate to more than one "parents" topic, that is, that polyhierarchies occurred. This happened in all those cases in which two identical descriptors were present at different levels in the original schemes. This type of relation had been dealt with using the symbol ▲ in the manual unification process. Figure 2 exemplifies that with the "terminology" topic for which distinct subordination could be observed.

The occurrence of polyhierarchies, allied to other factors that will be commented on later in this article, make it evident that merging the maps at the structural level results in a sum or union; that is, for those cases in which the same descriptor occurs in both maps, the two hierarchies in which they occur remain after merging.

By comparing the two schemes, one observes that the terminological variations, even when small, ex-

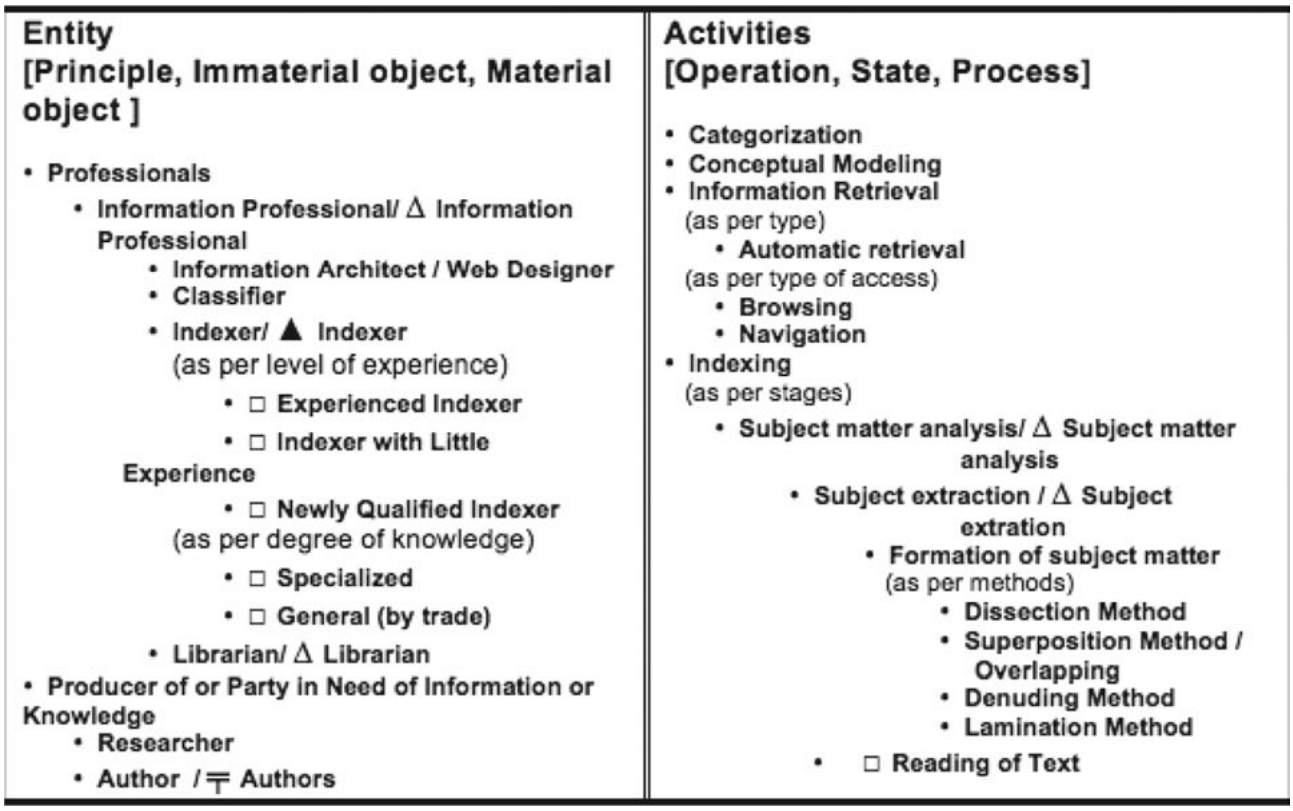

Figure 1. Scheme resulting from the manual unification.

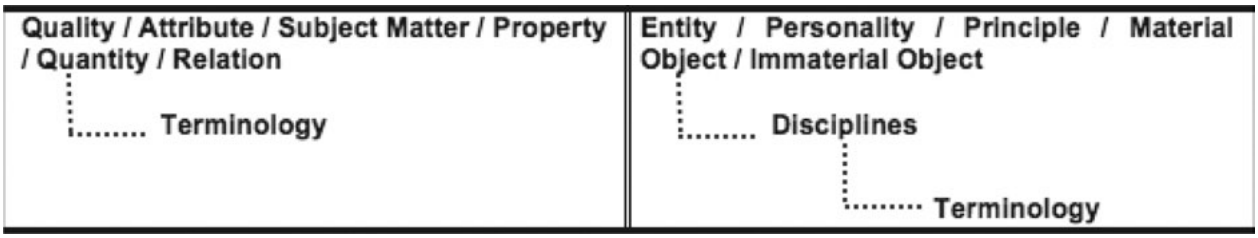

Figure 2. Example of polyhierarchy in the topic map

erted negative influence on the process of merging the maps, as among the characteristics of the topics used for the identification of equivalences are the names that are attributed to them. The "authors" and "author" topics, for instance, which in both maps represented the same subject, were not unified. This means that the relations identified by the ∓ and ■ symbols in the manually unified scheme did not receive the treatment expected in the automated merging. As to those topics that took part in relations identified by the symbol Δ (terminological and conceptual equivalence), they were duly unified.

It is important to note that in those cases in which two topics were unified into one, all properties of both topics such as names, associations, and, mainly, occurrences were kept in the resulting topic. Therefore, at the moment the user, when navigating through the map, reached a topic resulting from the unification of two topics (one from each map used in the input) the two relevant information resources were displayed.

Those cases that were treated with the ◊ symbol (terminological divergence, conceptual divergence only in relation to the specificity level, and structural equivalence) did not receive any type of treatment in the automated process, and were presented in the resulting scheme as non-related topics. Finally, the cases of simply structural equivalence, represented in the manual unification by the symbol □, were duly treated in the resulting merge topic map mainly thanks to the terminological compatibility between the large categories dealt with in the two schemes.

6.0 Final considerations and future developments

The topic map merge may be defined, within the scope of interoperability between vocabularies, as an automated terminological compatibilization, based on topic names; that is, two topics in different maps that have one of more common descriptors are unified into one single topic in the resulting map. It is in this topic that the sum or union of the characteristics of the original topics is deposited. Occasional variations in the utilization of descriptors may influence negatively the result of the process, even when it is with regard to small variations, such as divergence in the utilization of a descriptor in the singular or plural. Some future studies may propose the incorporation of algorithms for the mating of spelling verifications and standards, such as those currently used by Google, into topic map manipulation tools, thus providing for the treatment of small terminological variations or spelling mistakes. The structural compatibilization while merging takes place as a consequence of occasional unifications at the terminological level.

At the conceptual level, the proposal in the topic maps is to use the PSI concept. However, the identification of equivalences between one new topic and existing PSIs is a completely manual process, which requires intense intellectual effort on the part of the user responsible for the new map. That is, the utilization of PSIs, which is of fundamental importance if topic maps are to accomplish their goal of interoperability, is similar to the process of utilizing intermediary languages for the compatibilization of vocabularies.

Research into the compatibilization of vocabularies may come to contribute towards the development of new topic map interoperability techniques. The great majority of these studies, whose initial development dates back to 1960s, were focused mainly on its application in thesauri. It was only more recently that these studies have begun to be revisited with the goal of integrating other forms of information representation, such as ontologies, for example).

The topic maps generated from faceted schemes are extremely simple from the point of view of the resources offered by the maps (endless types of associations, names of topics, scope, etc.), due to a fragility of semantic explicitation of these schemes: types of limited relations, scope, and cardinality are not considered. In addition, polyhierarchy is not allowed, since the principles defended by Ranganathan establish that the ranking classes must be mutually exclusive; that is, no subject from the structure may belong to more than one class in the rank. Future developments may come to propose new subject matter analysis models based on the theory of facet analysis that may serve as the basis for the construction of more complete topic maps, under the semantic point of view, and more prepared for compatibilization through the utilization of scopes, for example. In addition to that, other proposals for the conversion of traditional IC classification schemes into topic maps may be developed, with the use of types of topics, associations, and roles that are different from those used in this project.

The occurrence of polyhierarchies in the map resulting from the automated merging proved to be efficient in the sense that it showed different conceptual definitions for one same term, the occurrence of which is clearly possible when we are working with information resources bearing different authorships. This fact acquires even greater relevance in those cases in which the content of the information resources that are being dealt with belong to knowledge areas whose terminology is not well defined. Such is the case with disciplines dealing with bibliographical classification (Satija, 2000), since, in these fields, authors will use terminologies that differ from those used in other fields. With no possibilities for polyhierarchies, there would be the risk of some term being housed under a subordination whose semantic value does not coincide with any of the occurrences that are likely to be accessed by the user.

For the same reasons, we must emphasize the importance of the possibility of attributing various names to the same topic. Once such equivalences are explicit, even when one any subject is referred to by different authors through distinct terms, the literary warrant remains. Most specifically in the cases of systems that favor context navigation to the detriment of other search techniques, the role literary warrant plays is even more important than that it plays in other systems that prioritize other search techniques. In the case of textual search, for example, the user warrant becomes essential.

In relation to the objective of the topic maps, which is to provide for contextualized navigation and offer greater flexibility for the representation of subjects and relationships, it is possible to say that the technology fulfills its role well. It supports a diversity of navigation interfaces (textual, hierarchical, in several forms of conceptual maps, or in various other graphic forms), with most of them permitting the user to efficiently access the location of any hypertext knots in relation to its global structure. Such efficiency is founded on mapping directly between the

structure represented by the map and the sources of information to which this structure refers. Additionally, there are no restrictions or limits for semantic representation; any idea, object or thought, as well as any type of relation they may have, may be represented by means of topic maps. The possibility of establishing limits for the application of certain subjects (scope) and the possibility of representing one same subject in different hierarchies (polyhierarchy) on in the same map contribute even further to its representation flexibility.

With regard to the semantic integration of heterogeneous schemes and the shared construction of conceptualizations, which are also considered objectives of the topic maps, we observed the confirmation of what some language and vocabulary compatibilization researchers had already observed: complexity of the compatibilization at the conceptual level, requiring a high level of manual interference. From that point of view, the topic map standard comes to simply play the role of supporting human analyses and decisions, since the automation of the process is concentrated at the terminological level.

With the evolution of an XTM standard, the development of languages for specific consultations and the possibility of implementing inference rules, new projects will be undertaken in an effort to reassess the level of automation of merge procedures.

Notes

1. Some examples of such restrictions are: "Every person was born somewhere" (considering a map approaching both the topics "person" and "place"); "An animal is always the offspring of another animal."
2. According to Maniez (1997) the words "integration," "harmonization," "reconciliation," and "concordance" are usually applied to indicate the concept of convergence, that is, the same concept of "compatibilization." However, Lancaster (1986) considers integration of vocabularies as a specific method of compatibilization, in which there is no conceptual analysis, that is, for a certain user request, all related occurrences (including word variations) and elements are shown as a result, without showing the relation or the context in which each one of them are present.
3. Polyhierarchy occurs when a specific term may be subordinate to more than one generic term. The term "geological chemistry," for example, can be subordinate both to chemistry and geology.
4. TMProc was developed by the Ontopia group (http://www.ontopia.net/) and was one of the first software applications developed for the manipulation of topic maps.

References

Ahmed, Kal, and Moore, Graham. 2006. An introduction to topic maps. *Architecture journal* July. Available http://msdn.microsoft.com/en-us/library/aa480048.aspx.

Campos, Maria Luiza de Almeida. 2005. A problemática da compatibilização terminológica e a integração de ontologias: o papel das definições conceituais. In *Anais do VI ENANCIB: Encontro Nacional de Pesquisa em Ciência da Informação, 28-30 Nov. 2005, Florianópolis*. Florianópolis: Universidade Federal de Santa Catarina.

Coates, Eric J. 1970. Switching languages for indexing. *Journal of documentation* 26:102-10.

Colmenero Ruiz, Maria Jesús. 2005. Introducción al modelo topic maps (ISO/IEC13250:2003). *Revista digital de biblioteconomia e ciência da informação* 3 n1. Available http://www.brapci.ufpr.br/documento.php?dd0=0000007483&dd1=0f592.

Dahlberg, Ingetraut. 1978. A referent-oriented, analytical concept theory for INTERCONCEPT. *International classification* 5: 142-51.

Dahlberg, Ingetraut. 1981.Toward establishment of compatibility between indexing languages. *International classification* 8: 86-91.

Doerr, Martin. 2001. Semantic problems of thesaurus mapping. *Journal of digital information* 1n8. Available http://journals.tdl.org/jodi/article/viewArticle/31/.

Garshol, Lars Marius. 2004. Metadata? thesauri? taxonomies? topic maps! making sense of it all. *Journal of information science* 30:378-91. Available http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html.

Garshol, Lars Marius. 2002. What are topic maps. *XML.com* September 11. Available http://www.xml.com/pub/a/2002/09/11/topicmaps.html.

Glushkov, Victor M., Skorokhod'ko, Ėduard Fedorovich, and Strongnii, A.A. 1978. Evaluation of the degree of compatibility of information retrieval languages of document retrieval systems. *Automatic documentation and mathematical linguistics* 12n1:18-26.

Guinchat, Claire, and Menou, Michel. 1994. *Introdução geral às ciências e técnicas da informação e documentação*. 2. ed. Brasília: IBICT.

Hudon, Michèle. 2004. Conceptual and lexical compatibility in thesauri used to describe and access moving image collections. In Julien, Heidi, and Thompson, Sharon eds., *Access to information: technologies, skills, and socio-political context, proceedings of the Congress of the Canadian Federation for the Humanities and Social Sciences, June 3-5, 2004, Winnipeg, Manitoba, University of Manitoba*. Available http://www.cais-acsi.ca /proceedings/2004/hudon_2004.pdf.

Lancaster, Frederic Wilfrid. 1986. *Vocabulary control for information retrieval.* Arlington, VA: Info Resources Press.

Librelotto, Giovani Rubert. 2005. "Topic maps: da sintaxe à semântica." PhD diss., Universidade do Minho. Available http://tede.ibict.br/tde_busca/arquivo.php? codArquivo=486.

Lima, Carlos Eduardo de, Fagundes, Fabiano. 2004. Utilização de mapas de tópicos no desenvolvimento de hiperdocumentos educacionais. In VI Encontro de estudantes de informatica do estado do tocantins – ENCOINFO 6, 4 e 5 de novembro de 2004, Palmas, Tocantins. Palmas, Tocantins: CEULP/ULBRA.

Lima, Gercina Ângela Borém de Oliveira. 2004. "Mapa hipertextual (MHTX): um modelo para organização hipertextual de documentos." PhD diss., Universidade Federal de Minas Gerais, Belo Horizonte.

Maniez, Jacques. 1997. Database merging and the compatibility of indexing languages. *Knowledge organization* 24: 213-24.

Naves, Madalena M. L. 2000. "Fatores interferentes no processo de análise de assunto: estudo de caso de indexadores." PhD diss., Universidade Federal de Minas Gerais, Belo Horizonte.

Neville, H. H. 1970. Feasibility study of a scheme for reconciling thesauri covering a commom subject. *Journal of documentation* 26: 313-36.

Newman, Simon M. 1965. *Information systems compatibility*. Washington, D.C.: Spartan Books.

Park, Jack, and Hunting, Sam. 2003. *XML topic maps*: creating and using topic maps for the web. Boston: Addison-Wesley.

Pepper, Steve. 2000. The TAO of topic maps: finding the way in the age of infoglut. In proceedings of the XML Europe conference, 2000, Paris. Available http://www.ontopia.net/topicmaps/materials/tao.html.

Pepper, Steve, and Moore, Graham. 2001. *XML topic maps (XTM) 1.0.*: Topicmaps.Org Specification. Available http://www.topicmaps.org/xtm/.

Rada, Roy, and Martin, Brian K. 1987. Augmenting thesauri for information systems. *ACM transactions on office information systems* 5n4: 378-92. Available http://dl.acm.org/citation.cfm?id=42246&dl=ACM&coll=DL&CFID=144583757&CFTOKEN=16357518.

Satija, Mohinder Partap. 2000. Library classification: an essay in terminology. *Knowledge organization* 27:221-29.

Silva, Marcel Ferrante. 2007. Estudo comparativo entre interfaces hipertextuais de soft-wares para a representação do conhecimento. *Ponto de acesso* 1n2:3-21.

Soergel, Dagobert. 1974. *Indexing languages and thesauri*: construction and maintenance. Los Angeles: Melville.

Soergel, Dagobert. 1982. Compatibility of vocabularies. In Riggs, Fred W. ed., Proceedings of the Conta Conference on Conceptual and Terminological Analysis in the Social Sciences, May 24-27, 1981, Bielefeld, FRG. Frankfurt: Indeks Verlag, pp. 209-23.

Tennis, Joseph T. 2001. Layers of meaning: disentangling subject access interoperability. In proceedings of the 12th ASIS SIG/CR Classification Research Workshop 2001, Washington, DC. Washington, DC: University of Washington.

Zhang, Xueying. 2006. Concept integration of document databases using different indexing languages. *Information processing & management* 42:121-35