tripleC 7(2): 309-322, 2009
ISSN 1726-670X
http://www.triple-c.at

Indexing languages in information Management, a promising future or an obsolete resource?

Jose Antonio Moreiro

Dept. Biblioteconomía y Documentación, Univ. Carlos III. Getafe (Spain) E-mail: jamore@bib.uc3m.es

Jorge Morato, Sonia Sanchez-Cuadrado, Anabel Fraga

Dept. Informática, Univ. Carlos III, Leganés (Spain) E-mail: {jmorato, ssanchec, afraga}@ie.inf.uc3m.es

Abstract: *Indexing languages have traditionally been an essential tool for organizing and retrieving documental information. The inclusion of indexing languages into the digital environment leads to new frontiers, but also new opportunities. This study shows the historical evolution of the indexing languages and its application in document management field. We analyze diverse trends for their digital use from two perspectives: their integration with other digital and linguistic resources, and the adjustment of them into the Web environment. Finally, there is an analysis of how these languages are used in the Web 2.0 and the incorporation of ontologies in the Semantic Web.*

Keywords: Indexing language, controlled vocabulary, Information representation, knowledge organization systems, Semantic Web, Topic Maps, Web 2.0

Acknowledgment: The authors wish to thank Rosa Macarro, Jose Maria Diaz Nafría for polishing the translation. This work was carried out within the framework of a research Project financed by the Spanish government (Ministerio de Educación y Ciencia, Secretaría de Estado de Universidades e Investigación, TIN 2007-67153).

Indexing languages are a key piece in the information management systems. These languages avoid the ambiguities of natural language using subsets of term called controlled vocabulary. The application of these languages supports the fair identification of main elements in the text. Also, they have a historical path, but digital media in heterogeneous environments multiply viewpoints and goals of final users when retrieving documents. This phenomenon entitles synchrony problems between the information analysis and the user expectative. As a solution in diverse digital environments, a tendency rejecting the application of controlled vocabularies in document descriptions appears. Moreover, new controlled languages demand for a place in the Web. In the following section, a historical evolution of indexing languages is showed, and after that the use of them in indexing description. Finally a reflection is done regarding its adaptation into the digital environment.

1. Categorization of Concepts in Symbolic Structures

Retrieval methods always support the task of finding a document in the stored documents. It happens with books, articles and web pages. In order to improve the information retrieval, the task of analyzing papers is needed; this process goes over two phases: *analysis* and *synthesis* of results.

During the *analysis* process, symbolic structures are identified in the semantic categories. At this point, Aristotle becomes a

first character; he considered predicates as main concepts pointed by the expression of

> *Substance, quantity, quality, relation, place, time, posture, vesture, activity, and passivity.*

This list has been reduced by Leibniz into five categories (Leibniz, 1966):

> *Substance, quantity, quality, relation, activity or passivity.*

Moreover, Kant proposes the pure concepts of understanding. Kant categorizes theses pure concept into four groups that are shown in Table 1.

Beyond substantive ideas, Porphyry (Plotinus's disciple) has the merit of having arranged the first semantic network by distributing universal predicates as a tree[1], showing graphically the relations between concepts regarding their genus, subtype and difference (Sowa, 2000; Moreiro, 2006). By these means, the existence of a hierarchic order between Aristotle categories is fixed, where genus is occupied by substance or composed by, descending in the scale in the order provided by *Genus and Species*. This order has arrived to us as a conceptual structure of taxonomies and thesauri, containing the source of the hierarchic disposition of their terms in its category relation. Each genus has as generic its immediate superior genus, for which it is species, at the same time that it acts as generic of inferior genus of immediate order. It means that a concept might be genus by relation of ideas, and species by its subordination to a general one.

Genus, with a supreme genus:

- *Top Term* or Macro-descriptor.
- With subordinates genus and species (*intermediaries*): *Middle Term* (Sub-macro-descriptor).

And *Species* (specific of different levels in a thesaurus):

- Species: Generic.
- Individuals: Specific.

Ramon Llull, following Porphyry categories more than Aristotle primitives, presented a semantic tree with seven parts, among them: structure, predicable and predicates or categories. In the first distinction –composed by *ens, substance, cors, animal and quaestio*–, he presented as fundamental methodology of his *Logica Nova* and, therefore, as universal mechanism of any communicative fact, the ten general rules of questioning[2]:

Utrum	Is it or not
Quid	What
De quo	Whom
Quare	Why
Quomodo	How
Ubi	Where
Quando	When
Quantum	How much/how many
Cum quo	With whom
Quale	Which

[1] Porphyry arranges Aristotle's categories as a tree (Arbor porphyriana) (Wildgen, 1994). In this tree is proposed the substance as a Summun Genus (general term), discerning between corporeal and incorporeal substance. The Corporal Substance is split into sensitive and insensitive. Finally, sensitive substance is subdivided in rational and irrational.

[2] Every logical reasoning must follow the hermeneutic decametre (Llull, R., 1970).

Table 1: Kant Categories (Kant, 2002)

Quantity	*Quality*	*Relation*	*Modality*
Unity	Reality	Inherence and Subsistence (substance and accident)	Possibility—Impossibility
Plurality	Negation	Causality and Dependence (cause and effect)	Existence—Non-existence
Totality	Limitation	Community (reciprocity between agent and patient)	Necessity—Contingency

The Port-Royal movement introduces essential elements, situating the logical concepts of *Definition* and *Division*: the first to explain the *quidditas* of something, as the way in which the meaning of some descriptors are specified in thesauri by using Scope Notes. The *Division* clarifies the difference between species, analyzing the genus by means of the differences[3].

The organization of concepts built by Rhetoric has arrived to us[4]. Whenever a proposal has been elaborated for categorizing text concepts, then relation between concepts appears as a fundamental way of thinking. These concepts, and the semantic relationships among them, show the knowledge usually represented in thesauri, taxonomies and ontologies. In fact, diverse Lulian's reasonings are compiled in Semantic Networks and Artificial Intelligence (Boden, 1994).

Thus, general mechanisms of reasoning proceed establishing relations between concepts. This behaviour has been transmitted, in a peculiar way, to the elements that integrate indexing languages:

- Terms integrated in the same category.
- Difference between species.
- Division or analysis of genus by differences.
- Scope Notes, for explaining meaning of terms.

Thesauri fundamentals can be even found in the theory of Derrida's Deconstruction, by arguing that a linguistic sign may be repeated.[5] In order to avoid some Nietzsche's contingencies, such as the possibility of saying nothing, or the danger of name misappropriation, Thesauri agree to call univocally every concept by a term, avoiding ambiguity and giving guarantee of concept understanding independently of situations. One step forward was given by *Topic Maps*, establishing relations between diverse texts acting as Metaindexes.

1.1. Documents' Content Representation

Documents, in order to be informative, must be communicated. That is the reason why their ideas must be ordered for readers, to be understood.[6] Hence, the determinant function of macrostructures in texts is distinguished. Dual composition of significant/signification offers a parallelism in texts, because sentences and phrases result from the union of the expressive plane (syntactic) with the content one (semantic, conceptual), to which

[3] Llull proposal influenced Leibniz and Descartes, fathers of Port-Royal Logic, which therefore agree with Porphyry and Llull considering five predicates instead of four, and including species between universal ideas: genus, species, differences, properties and accidents (Arnaud & Nicole, 1987).

[4] In medieval and ancient times, Poetic and Rhetoric studied conceptual principles of discourse. Poetic focused on syntactic-structural organizations of literary texts while Rhetoric focused on non literary texts (Wildgen, 1994).

[5] The reader questions himself/herself; he/she is co-part of the writing, by deconstruction. So, as strategy of writing and reading, it is read and written by a splited gesture (repeated). (Derrida, 1975)

[6] The schema follows Hjelmslev ideas, the linguistic sign from four levels disposed in two correlated pairs. In one side content-substance, and in the other side the content-form (Hjelmslev, 1986).

the relation between author and readers, and both with the message (pragmatic) must be added. None of these elements could be forgotten whenever a semantic analysis is done, even if the only interest was to indentify the contained essential concepts.

Among the theoretical contributions in the analysis and representation of contents, coming from linguistics disciplines:

- It is relevant the **conceptualization** built from ancient times by Rhetoric, having continuity in our days through **Text Linguistics**[7].
- It is also important the participation of **Language Technologies with syntactic and semantic analyzers** in any automatic indexing program.
- Even **Lexicographic** intervention is determinant.

1.2. Indexing Languages

Indexing languages show nowadays a similar complexity to that of networks and informatics systems through which the information to be represented flows. Thus, its comprehension must be taken from conceptual basis coming from diverse fields like Formal Logic, Statistics, Rhetoric, Linguistics, Semiotics or Lexicography.

What was common until some decades ago is still valid. Nevertheless, indexing languages must be currently translated to the digital context. In figure 1 these languages are shown.

Zeng (2005) establishes that indexing languages are, in fact, knowledge organization systems. We can classify these systems, regarding the level of structure complexity and the simplicity of implementation, in four classes (Figure 1):

- **Keywords and folksonomies**: Actually, they are not languages, but not-controlled vocabularies. They have a relevant role in the social Web.
- **Wordlists**: Glossaries and gazetteers.
- **Faceted classifications** and **library categorization schemas**.
- **Relationship groups**: based on associations between concepts like thesauri, Topic Maps, and ontologies.

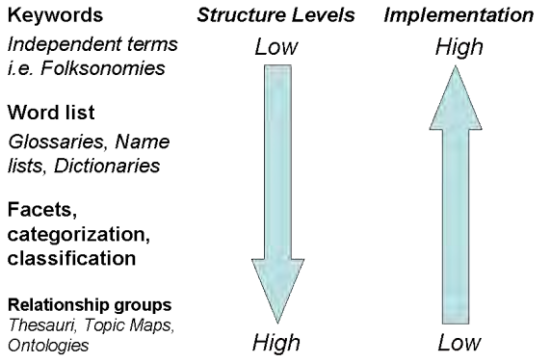

Figure 1: Indexing languages used in retrieval systems

2. Main Concepts and Their Macrostructural Organization

The varied structures and meaning elements integrating documents need a global meaning for order and enclosure. Words and sentences are basic components of texts which integrate into other elements of higher category. The semantic levels organize each text in essential parts for its understanding, depending on macrostructures.

Macrostructures represent information contained in part of a discourse (partial macrostructures) or the whole discourse (global or general macrostructure). It is the deepest meaning structure, representing the general content of texts (theme and subject) and linking with the rest of the semantic elements:

[7] The relationship between Text Linguistic and Information Science is argued in Moreiro (1993).

Table 2: Speech/Discourse of Semantics Units

Meaning structure type	*Text semantic Level*
Superficial structure	Microstructures Sentences and paragraph structures
Intermediate structures	Partial Macrostructures Secondary and partial structure
	Superstructure Partial macrostructure order
Global structures	General Macrostructure Global semantic structure

Without a general macrostructure, the coherence of the text would be superficial and lineal (Dijk, 1980). A deconstructive position is adopted to analyse the documents, counter to authors writing. Necessarily, the syntactic, semantic and pragmatic role played by the *superconcept* has to be recognised from the very beginning. This superconcept, named *macrostructure* (**Mg**), represents the meaning of essential concepts. Then, relevant information could gradually be identified, since the main semantic aspects of the core idea must be expanded to partial and secondary macrostructures. Partial macrostructures represent concepts only important in some parts of the text; thus, they are considered as *submacrostructure* or *secondary macrostructure* (*sM*). Depending on the *sM*, another macrostructure even more *partial* (**pm**) coincide with text fragments of minor rank. Finally, there are *microstructures* (mi), with superficial meaning because of local links in a phrase or sentence. They have a low significance level to understand the content.

Partial macrostructures arrange document components and facilitate the access to contents. These macrostructures are essential to analyze and understand discourse, because of linking subsets of meaning for each discourse division. Since they are intermediate level structures, they play the role of secondary macrostructures, reaching in a descending path to the partial or modular macrostructures (pm). This sequence of composed structures has been formalized, as shown in the common convention of chapters and paragraphs in books, or movie sequences.

Thus, every document model has adopted a specific archetypical device. Not only a global semantic structure is presented, it also generates a global schema within its text development, known as *superstructure*, in charge of:

- Formal organization of intermediate structures.
- Coherent succession of partial macrostructures existing for singular function.

Its place in the analysis of contents arises from the fact that many documents follow a specific schema with functions to accomplish, which is a consequence of the fact that the superstructure serves as auxiliary for creating a document and then for consulting it. The superstructure represents the course followed by the content in its progress from the general macrostructure to each of the microstructures, which may be schematized in a series of categories hierarchically ordered. It also allows signaling which parts of the document contain the universal information that can be represented.

Each text model is organized in an elemental superstructure, which is transformed depending on its specific global macrostructure.[8] This is the reason why text schemata are similar for any juridical norm, commercial letter, TV news, administrative or scientific documents.

Discourses have also microstructures, whose limited semantic relevance only unifies a sentence. In this case, expression is concreted by *superficial structures* of low importance for library science experts because of the partial message being

[8] Drop (1987) defines it as a schema (although not conventional) of a succession of (macro-) "verbal" acts, in order to find a "thread" for the text. Scientific documents follow a canonical division in its organization: Parts, Chapters, Sections and Paragraphs. The archetypical presentation of empiric researches is shown as: Problem hypothesis, research method, results and discussion, conclusions or recommendations. (Bernárdez, 1995); Trujillo, 2002).

contained.[9] At the moment of building a text, authors proceed combining discrete microstructure units. Also readers might recognize the content of the texts by following microstructures, but very unusually they explain the substance of the text as in the case of being located in the introduction, other main paragraphs of a work, or in moral of stories.

The macrostructural levels of a text are considered when deepness is analysed. Deepness influences the type of results in the content analysis:

- In the abstract, informative and indicative models are distinguished for representing more or less partial macrostructures.
- In the combinatory index, more or less terms will be used, if global or also partial structure is represented.

In specific domains, indexing every concept in a document might not be useful for users, since query needs could be deceived, due to the fact that paying attention to microstructures entail non relevant concept representation. They do represent neither the global meaning of a text (general macrostructure) nor of its parts (partial macrostructures). Microstructures could have low relevance for retrieval, although full text indexing has great popularity among internet users. Internet search engines need this short of indexing due to the fact that:

- The Web has a heterogeneous nature, and the users have different motivations to search. Therefore, a quite specific search performs better improving recall ratio than precision ratio.
- Search engines decrease the weight of microstructures, by means of heuristic rules and word frequencies (i.e. a single occurrence of a word in the last paragraph of a document has less importance than a word that appears fifty times in the beginning of the text).

[9] Superficial structures support the linguistic framework within the limits of a sentence, formed result of rhetoric *elocutio*: they enable the linear concatenation of sentences in macrostructures (Albadalejo, 1989).

An example might clarify this issue: if a user searches the frequency of a term “t of student” in biomedical literature, he/she will have to search for the term independently of the role it has in the document. What makes the term more relevant is the change in the purposes of the user with respect to the document. In the hospital documentation centre, the macrostructure of documents becomes typical for medical domain. Just in this domain, examples are available everywhere, showing sites with controlled indexed languages for pointing macrostructures[10].

Since the 16th century B.C, in Babylon libraries, a division of sections for information retrieval has been done, ordered by categories. About 300 years B.C. Callimachus divided the Alexandria Museum in 127 discipline sets, the *pinakes*, in order to identify and organize the paper rolls following the knowledge classification provided by Aristotle (Millares Carlo, 1971). This approach is similar to that proposed by Benveniste (Benveniste, 1989), who suggested the use of the own scientific-technical terminology for concept representation and specialized knowledge transmission.

Thus, Lexicography and Semiotics has always participated in formalizations assuring social knowledge flow. From a contemporaneous perspective, two antagonist books appeared in the United States of America in 1876: the *Decimal classification* by Melvin Dewey (1979), which marked the path to be followed by classification systems, pre-coordinated and hierarchical structures; and the *Rules for a dictionary catalogue* by Charles A. Cutter (1962), who considered subject heading lists and somehow also controlled vocabularies. The Cutter theory has prevailed due to:

- Pre-coordinate character.
- Associative structure.
- Control of specialized vocabulary for indexing concepts.

[10] An example is given by the Medical Subject Heading in Medline - MeSH (PubMed <www.ncbi.nlm.nih.gov/pubmed/>)

- Improved usability.
- It is coupled with controlled vocabularies.

While Dewey's classification consists of:

- Analysis of human knowledge.
- Focus into decimal coded divisions.
- A valid universal expression that offers a general vision of knowledge.
- Achievement of a gradual domain sequence starting from the most global knowledge classification level (Bacon, 1980), going through intermediate levels suggested in the century XIX, and finally arriving to specific levels (Cutter, 1972).

The classification has continuity in La Fontaine and Otlet, in a positivist point of view.[11] One of their main tasks was the launching of the *International Institute of Bibliography* aiming to develop the *Universal Bibliographic Repertory* (RBU), whose development required the use of the Universal Decimal Classification (UDC), since, without it, international cooperation was not possible (Levie, 2006). The UDC –adaptation of classification conceived by Dewey– related concepts of the repertory by means of hierarchy, similarity or difference (association). Thus, as well as taxonomy, association was considered, which besides has been a constant technique in Library Science, as it was previously mentioned concerning the indexing of books.

The UDC was a way to organize information for later effective recovery. Its content organization achieved the knowledge management demanded by society. Nevertheless its classification system –as it happens with nomenclatures– became restrictive for desired user operations.

The work of information organization for retrieval does not achieve its objectives if it is limited to the use of references –characteristic of classical lexicography–, as it happens with universes:

- Not constituting a formal domain clearly limited,
- In transformation,
- Characterized by multiple interdisciplinary relations,
- With low levels of order, or
- Not analyzed from agreed organization parameters.

Hence, terms achieve a double function in knowledge transmission (Cabré, 1999): the denominative function, and the conceptual one, although with different levels and modes, and in diverse situations.

The definition of concept in traditional terminology is accepted as restrictive. For that reason, the social cognitive conceptualization has been searching for a real description of the meaning of terms as they appear in texts. This fact breaks the centralized vision intended by standards (Temmermman, 1999), because the content has passed to be limited by the context, where the term is inserted. A concept is neither universal, nor immutable. Furthermore, it is elaborated by the knowledge of the world, and by linguistic-semantic understanding, which enables exact sense representations among the concepts of a text (Moreiro, 1993).

Another decisive fact was the massive information gathering during the Second World War; overflowing the methods for transmitting and accessing results in research. Vannevar Bush, Director of the *Office for Scientific Research and Development* with President Roosevelt, wrote about its experience in the paper called "*As we may think"*, where he stated the arguments for understanding what decades later have been known as *Information Society*. He stressed the importance of communication and availability of information for generating new knowledge, considering the procedures and instruments that could be useful for knowledge control and organization. Bush understood that sequential structure of documents –reflection of sequential oral discourse- caused that the alphabetical or numerical taxonomies, arranging concepts in class-subclass, were not suitable for an adequate processing of high amounts of information, since "the human brain does not

[11] In Science Classification, Comte postulated that sciences present complex and interdependent relationships (Ducasee, 1950).

work in this way, but it works by associations" (Bush, 1945).

The main problem, according to Bush, was the inadequate manner for storing and classifying information. Hence he imagined a more effective system for processing, the MEMEX:

- Anticipating the importance of retrieval by logic combinations of document subjects.
- Overcoming taxonomic structure.
- Preserving associations between concepts, therefore, imitating the mode people think in order to achieve "association indexing" (Bush, 1987).
- Beyond information storing, grouping and linking documents in hypertexts, as an alternative to knowledge linear retrieval.

Users could follow multiple trajectories, as a new text reality (hypertext), and also as a new way of reading and writing[12]. The proposal of MEMEX introduced more flexibility. It is evident that traditional categories of Library Science are not sufficient for content treatment of documents, because universal classification does not help much to the effective flow of contents. Bush Bush initiated the use of associative indexes and predicted the use of databases, hypertext and hypermedia (Roberts, 1976; Buckland, 1992).

Other foundations are rooted at the development of personal computers and documental techniques fostering databases development and the usage of coordinate languages. Thus a relation between Lexicography and Indexing languages was established by the proposal of norms ISO 704-2000 and ISO 1087/1-2000. These classic standards have contributed to the development of methodologies for indexing languages. Among these, thesaurus has become a reference, since it represents the concepts of a specialised domain in a normalised way, using univocal terms, which are structured following logic-semantic principles:

- Each descriptor is situated in a sense-giving context, by relations between terms.
- The set of expressions, fixing a specific concept, are exhibited.
- The conceptual proximity between terms and descriptors is shown.

Indexing languages lack of popularity is due to the absence of the desired double articulation, towards users and back from users; thus losing connexion to reality (Gonzalo & Yebra, García, 2004). This explains why –as shown in Figure 2– the higher the complexity of the language is, the lower the proximity and usability from a user point of view. (Morato et al., 2008). Undoubtedly, lightly formalized Knowledge Organization Systems (KOS) (Light weight ontologies) improve the usability, management and understanding on behalf of the user, which is obvious if we observe the Web 2.0 tags popularity (also called folksonomies) (Gruber, 1993) for describing multimedia resources of the Invisible Web. The Web 2.0 was developed as a natural evolution of the Web 1.0. Social Web, or Web 2.0, is a more usable Web employing a minimum complex indexing language (see Figure 3).

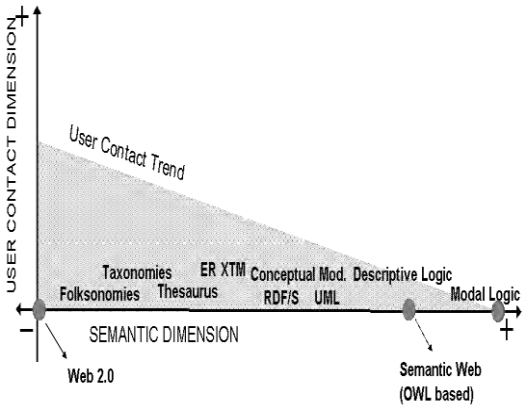

Figure 2: Semantic complex and usability relation

The Web 2.0 (O'Reilly, 2005) offers the possibility of achieving everything at the same time. The advantages offered are quite clear in the user interaction, because it is greater than before; as shown by some social

[12] Some theorists as Landow (1995) have lead this trend to the extreme of disappearing authorship and traditional texts, which is replaced by the absolute success of readers.

networks like: Orkut, LinkedIn and Multiply, platforms like Windows "live", and Google, RSS, blogs, photofloods, folksonomies, P2P platforms, and so on. But Web 2.0 is also a space for free applications: Free software (Moreiro, 2006) and free files offering applications and open documents (Tramullas, 2005; Stallman, 2004). So it could be stated that a definite profile of Web 2.0 is its participative character, a free space for collaboration and communication with a different response for indexing languages, higher association richness, better adapted to change by means of a higher proximity to user modes and needs.

Although from a user perspective the Web 2.0 is more intuitive, its evolution should integrate semantic applications through more powerful and more formalized indexing languages (for example, ontologies based on XML[13]). These applications constitute the Semantic Web.

The convergence of both ideals, Semantic Web and Web 2.0, opens up three possible trends of the Web (Figure 3). The Evolution of Web 2.0 to Semantic Web; the Convergence, with or without W3C intervention for a hybrid Web; or the Coexistence, where Web 2.0 is centred in the description and sharing of low value resources, and Semantic Web is specialized in critical resources.

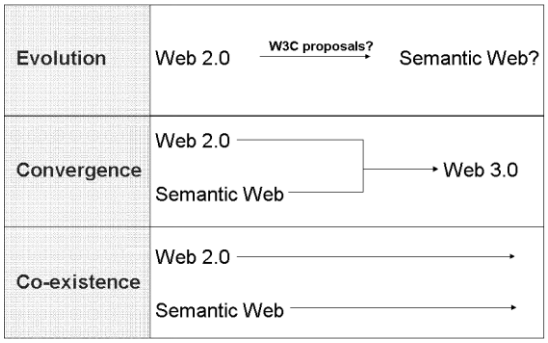

Figure 3: Possible evolutions for the Web

Internet is composed by natural language documents from heterogeneous fields. The indexing languages will accomplish its communicative, interpretative and retrieval functions, just assigning content metadata by clarifying links. Internet shows linguistic problems because of its ambiguity. This is why terminology is still a reference whenever words are located in contexts with different possible interpretations.

The main relation between Indexing Languages and Lexicography is originated by ISO 704-2000 and ISO 1087/1-2000 standards, due to its contribution for building indexing language methodologies. These standards grant a more complete ending than ISO 2788 for thesaurus construction (ISO 2788; ISO 50-106-90) due to:

- A clearer comprehension of Concepts.
- Exhibited relations between concepts.
- Facilities to know the configuration of conceptual systems.

New norms, like ISO standard for *Topic Maps*, merge the use of indexing languages (ISO/IEC JTC 1/SC34. 2004) with digital media, opening issues which were previously strictly regulated.

3. Information representation in current indexing languages

The name of *Conceptual Thesaurus* is founded on the notion of matter (in its conceptual sense) as substance of texts. It gathers terms and concepts by their similarity of sense within user context, and it is characterized by the richness of its associative relations (Maniez, 1993). Besides, it can be understood as a semantic network of concepts in which every node contains a unique semantic concept that may have various associated descriptors, which can also be identified in the network of related descriptors by means of the typical relation of thesauri: preferential, hierarchical or associative. With respect to traditional thesauri, they present the following new features:

- All non-empty words available in database are listed;

[13] Examples of Thesauri in XML are *Topic Maps* or Light Ontologies. Unfortunately, XML represents only syntax and it does not show a formal model for representing knowledge. This lack is overcome, however, by RDF language.

- Colloquial terms, even variations, and known truncated terms are taken into account.
- Defining notes dissipate possible use, doubts are included
- The equivalency of available terms is reasoned
- The associative relations are included, even with non descriptors

Examples of how these thesauri had evolved are MeSH and WordNet (Morato, Llorens & Marzal, 2004).

In its operation, the conceptual thesauri take advantage of the design of ontologies, which allow distinguishing synonyms, suppressing the homonyms and inducting associative relations between descriptors. Ontology for a knowledge base should embrace different types of documents, conceptual descriptions, relation among documents (quotes), and links to different scientific problems, as well as index, bibliographic descriptions, thesauri, sorting codes and terminological information. Its application must include a complete domain vision of the structure and terminology, which should facilitate relevant retrievals.

The communicative demands of the multimedia hypertexts has forced thesauri to be opened to new relations, higher in number and improving its representation, accuracy and efficiency. We may assert that the Internet and its hypertext linking offer of documents compelled to distinguish representation from documental contents, evolving from finding solutions firstly in digital thesauri to verbal and expanded thesauri, improving its expression through XML (as in topic Maps) or improving its visualizations using technologies like Flash.

Another proposal for improving thesaurus is the inclusion of verbs complementing static traditional thesaurus of substantives. This thesaurus is called *verb thesaurus* (Moreiro, 2000). This task might be facilitated by merging the thesaurus of substantives with verb classifications from the linguistic field (Levin, 1993). The use of verbal descriptors gain multiple advantages as the possibility for indexing audiovisuals by gerunds, identification of verbal functional associations more adaptable for specific domains, the possibility of showing existing relations between concepts using natural language facilities (verbal categories as relations that can be faceted), and conceptual disambiguation.

The trend toward increasing relations is common to similar technologies in diverse fields like Software Engineering, e.g. UML. This richer concern in relations may be exemplified by: *aggregation* relationship, in which the disappearing of the whole does not entail the disappearing of its parts; *composition* relationships, in which the parts disappear when the whole does; *multiplicity* information (i.e., how many objects may interact within the same relationship); relationship *direction*; or relationship *typification*.

The approach of verbal integration comes from pedagogy, where relations built by the use of verbs are called conceptual maps (*concept maps*). It could be considered as a precursor for WordNet lexical database, an interdisciplinary semantic network in English for conceptual disambiguation, using verbs (Moldovan, 2001) and offering equivalence and hierarchy relations between different grammatical categories (Green, 1995).

The automatic construction of thesauri by means of the nearest occurrence of nouns in verbal structures follows a process starting with the analysis of relevant documents for vocabulary extraction (glossary, dictionaries, etc). Afterwards, it is depurated by hand extracting descriptors, which are used for indexing documents (handbooks, standards, and so on). In this stage phrases are stored if one or more descriptors of the thesauri appear in their nominal Subject Syntagma; and one or more descriptors in their Verbal Syntagma. Then, dynamic concepts are grouped, classified and related. Finally, the result obtained is reviewed by hand.

An innovative case in the information representation field arises from *Conceptual Maps for Navigation* through semantic nets. Its study originated a need for building indexes for subjects. The semantic networks are a common method for representing knowledge in Artificial Intelligence field, where

the achievement of communication between people and computers is searched.

Conceptual maps offer relationship networks richer than thesauri. Their support to navigation is more natural due to its node-link-node structure. The use of these maps eases the development of mechanisms for representing and retrieving, because relations between concepts are chosen following needs and user expectative. It is a technique for representing knowledge in cognitive graphics, which gave rise to the well-known standard "*Topic Maps"*: a document or a set of documents SGML or XML interrelated in a multidimensional space where each node is a *topic.*[14]

A *topic* is a term that represents a concept or idea, whose characteristics are: names, occurrences, role associations (Moreiro, Sanchez-Cuadrado & Morato, 2006). The relations are tagged as verbs in structure: Topic-verb-topic, thus an association is a verbal link between two or more topics, enriching the net of relationships. A *topic* must have a *basename*, as a necessary element representing the common way to mention the topic; it may also contain *alternative names*, as *display names* (shown to final users), *sort name* (alphabetical order if a list is required).

Topic Maps present undoubtedly advantages, as the optimization of conceptual maps or the merging of vocabularies hierarchical or not. Moreover, it is an intuitive ISO standard for creation and interpretation. Together with RDF/OWL and UML, it is one of the most extended languages in the Semantic Web, suited for the development of sites and the extension of searches. On the contrary, it has the disadvantages of lacking inference, rules, axioms, or flexibility to become adapted in different contexts.

Ontology is a shared conceptualization, explicit and formal for a domain (Gruber, 1993); it is composed by terms and relations in a domain, with combinatory rules, as well as terms and relations for extending the vocabulary (Neches et al., 1991). Studer definition is important for stressing the exact meaning of each component: *conceptualization* (abstract model of real phenomena with relevant concepts), *explicit* (concepts, types and restrictions are defined explicitly); *formal* (legible by a machine); and *shared* (with agreed knowledge accepted by a community) (Studer, Benjamins & Fensel, 1998). Therefore, ontology specifies a concrete viewpoint, reflected in the concepts depending on the language used for its description.

No matter what type of ontologies, vocabulary always appears in representation. In the *KR (Knowledge Representation) Ontologies*, class names, relations and functions are the elements expressing the knowledge; in *Common Ontologies*, shared experiences are represented by vocabularies of things, events, time, space (e.g. a metric system ontology); in *Top Level Ontologies*, general concepts for anchoring root terms with other ontologies are used; in *Upper-Level Ontologies*, vocabulary and relations are included; *Task Ontologies* employ the needed vocabulary for each task: verbs, adjectives, names; and finally in *Domain Ontologies*, the vocabulary may be reused in some domains but not in others.

Ontology represents a cognitive organization developing a system of knowledge organization. Nevertheless, one of the main problems to represent knowledge is the agreement of what should be represented. For that purpose, from different disciplines (Library Science and Documentation, Artificial Intelligence, Software Engineering, Linguistics, Ontological Engineering, etc.) several representation models have been proposed (Sanchez-Cuadrado, Morato, Palacios, Llorens & Moreiro, 2007).

The required degree of semantic representation and its objective determines the models and languages to use whenever a knowledge system is built, taking into account those within the ontology spectrum.

The ontology, as system of knowledge organization, tries to represent generic and specific information exhaustively. Ontologies can be configured following various knowledge modelling techniques and can be built trough different formal languages. In many cases the languages for models of knowledge representation entail a complete paradigm and a support language.

[14] ISO/IEC 13250: 2000. SGML-Topic Maps.

4. Final considerations

An indexing language does not only synthesize relevant information, at the same time it works as a conceptual device for domain organization. Therefore, it can work as an axis of cooperation –communication engine- and of ability –engine for knowledge and information production-. It is considered that indexing languages grant the improvement, selection, processing and assimilation of available information.

The "new" thesauri improve, in many ways, the association of terms: hierarchical, associative, formal, conceptual, referential, explanative, and so on; and enable many different forms of representing concepts, from a strict sequential point of view to a combination of sequential relationships; several kinds of taxonomical or associative representations (Rodriguez Bravo, 2002). They even widely overcome the constraint to represent concepts just with substantives.

In Information Systems, the descriptor, considered as a unit of information in the indexing languages, is the axis around which indexing and content retrieval spins within knowledge organization of a specialized domain.

Indexing languages represent the key for information management systems in order to reduce the ambiguity of natural language. The objective identification of the main elements of a text is an outcome of valid historical traditions and the use of a diversity of digital environments for dissimilar types of documental recovery. In the Web, controlled languages must coexist with indexing and free-searching systems.

References

Albadalejo, T. (1989). *Retórica*. Madrid: Sintesis.

Aristóteles. (1982). *Tratados de Lógica (El Organon)* (M. Mart, Candel Sanín). Gredos.

Arnaud, A., & Nicole, P. (1987). *La lógica o el arte de pensar*. Madrid: Alfaguara.

Bacon, F. (1980). *Instauration Magna. Novum Organum*. Mexico Porrua: Nueva Atlántida.

Benveniste, E. (1989). *Problemas de lingüística general II*. Madrid: Siglo XXI.

Bernárdez, E. (1995). *Teoría y epistemología del texto*. Madrid: Cátedra.

Boden, M. (1994). *Filosofía de la Inteligencia Artificial*. México: Fondo de Cultura Económica.

Buckland, M. (1992). Emanuel Goldberg, electronic document retrieval, and Vannevar Bush's Memex. *Journal of the American Society for Information Science*, *3* (4), 284-294.

Bush, V. (1945). As we may think. *Atlantic Monthly*, *176*, 101-108.

Bush, V. (1987). MEME revisited. *reprinted in Evolution of an information society* (pp.179-191). London: ASLIB.

Cabré, M. T. (1999). *La terminología. Representación y comunicación*. Barcelona: IULA.

Cutter, B. A. (1972). *Dewey Decimal Classification*, in Encyclopedia of Library and Information Science. New York: Marcel Dekker. vol. VII.

Cutter, Ch. (1962). *Rules for a dictionary catalog*. 4th ed. London: Chaucer House.

Derrida, J. (1975). *La diseminación*. Traducción, J. Martín Arancibia. Madrid: Fundamentos.

Dewey, M. (1979). *Decimal classification and relative index*. 19th ed. Albany: Forest Press.

Dijk, T. A. Van. (1980). *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognitio*n. Hillsdale, NJ: Lawrence Erlbaum.

Drop, W. (1987). Planificación de textos con ayuda de modelos textuales, in Bernárdez, E. (comp.). *Lingüística del texto*. Madrid: Arco Libros.

Ducasee, P. (1950). La synthèse positiviste: Comte et Spencer en *Revue de Synthèse*, *26*, 154 163.

Gonzalo, C., & Yebra, García, V. (2004). *Manual de documentación y terminología para la traducción especializada*. Madrid: Arco/Libros.

Green, R. (1995). The Expression of Conceptual Syntagmatic Relationships: a Comparative Survey. *Journal of Documentation*, *51* (4), 315-338.

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

Hjelmslev, L. (1986). *Prolegómenos a una teoría del lenguaje*. Madrid: Gredos.

ISO /IEC JTC 1/SC34. (2004). Information Technology - Document Description and Processing Languages. SGML-Topic Maps http://www.topicmaps.com/content/resources\iso13250/iso13250-2000-fcd.htm

ISO 1087/1-2000. (2000). Terminology work. Vocabulary. Part 1: Theory and application. ISO Standards, TC 37/SC 1. Genève: ISO.

ISO 2788: 1986 (1986). Documentation. Guidelines for the establishment and development of monolingual thesauri. ISO Standards, TC 46/SC 9. Genève.

ISO 704-2000 (2000). Terminology work - Principles and methods. ISO Standards, TC 37/SC 1. Genève: ISO; Documentation. Guidelines for the establishment and development of monolingual thesauri. ISO Standards, TC 46/SC 9. ISO 1087/1-2000. Terminology work. Vocabulary. Part 1: Theory and application. ISO Standards, TC 37/SC 1. Genève: ISO.

ISO/IEC 13250: 2000. SGML-Topic Maps.

Moreiro, J.A.; Sánchez-Cuadrado, S.; Morato, J. (2003). Panorámica y tendencias en topic maps. Hipertext.net, 1, <http://www.hipertext.net> [Consulted: 17/05/2009]. ISSN 1695-5498.

Kant. E. (2002) *Crítica de la razón pura*. *Translation by* José del Perojo & José Rovira Armengol. 2 vol. Barcelona: Ediciones Folio.

Landow, G. P. (1995). *Hipertexto: la convergencia de la teoría crítica contemporánea y la tecnología*. Barcelona: Paidós.

Leibniz, G W. (1996). De Synthesi et Analysi universali seu Arte inveniendi et judicandi incluido. *Hauptschriften zur Grundlegung der Philosophie* v. I. Hamburg: Meiner.

Levie, F. (2006). *L'Homme qui voulait classer le monde, Paul Otlet et le Mundaneum*. Bruxelles: Les Impressions Nouvelles.

Levin, B. (1993). *English Verb Classes and Alternations: A preliminary Investigation*. Chicago: The University of Chicago Press.

Llull, R. (1970). *Ars generalis ultima. Mallorca*: *1645*. Reprint, Frankfurt: Minerva.

Maniez, J. (1993). *Los lenguajes Documentales y de clasificación: Concepción, Construcción y utilización en los sistemas documentales*. Madrid: Pirámide.

Millares Carlo, A. (1971). *Introducción a la historia del libro y de las bibliotecas*. México: FCE.

Moldovan, D. (2001). Question Answering Systems in Knowledge Management. *IEEE Intelligent Systems*, *16(6)*, 90-92.

Morato, J., Llorens, J., & Marzal, M. A. (2004). WordNet Applications, in: Sojka, P., Pala, K., Smrz, P., Fellbaum, C., & Vossen, P., *GWC 2004 Global Wordnet Conference*, pp. 270-278. Brno: Masaryk University.

Morato, J., Sanchez-Cuadrado, S., Fraga, A., & Moreno-Pelayo, V. (2008). Hacia una web semántica social. *El Profesional de la Información*, *17(1)*, 78-85.

Moreiro González, J. A. (1993). *Aplicación de las Ciencias del texto al Resumen Documental*. Madrid: Universidad Carlos III de Madrid - BOE: 45.

Moreiro González, JA, Morato, J., Llorens, J., Marzal, M., Beltrán, P., & Vianello, M. (2000). Desarrollo automático de un tesauro de verbos para entornos de información dinámica. On CD-ROM: Brasilia: ANCIB. pp.359.

Moreiro González, J. A.; Rodríguez Barquín, B; García Martul, D.; Y Pinto, A. L. (2006). Bibliotecas Digitales y Open Source Software. *Informação & Sociedade*: Estudos, *16(1)*, 9-22.

Moreiro, J. A, et al. (2006). Categorización de los conceptos en el análisis de contenido: su señalamiento desde la Retórica clásica hasta los *Topic Map*, in *Investigación Bibliotecológica*, *20(40)*, 49.

Neches, R., Fikes, R., Finin, T., Gruber, T. R., Patil, R., Senator, T., et al. (1991). Enabling technology for knowledge sharing. *AI Magazine*, *12* (3), 36-56.

O'Reilly, T. (2005). What Is Web 2.0. Design Patterns and Business models for the Next Generation of Software. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html [Consulted: 29/10/2008].

Roberts, N. (1976). The pre history of the information retrieval thesaurus. *Journal of Documentation*, *40* (4), pp. 271-285.

Rodriguez Bravo, B. (2002). *El documento: entre la tradición y la renovación*. Gijón: Trea.

Sanchez-Cuadrado, S., Morato, J., Palacios, V., Llorens, J., & Moreiro, J. A. (2007). De repente, ¿todos hablamos de ontologías?. *El Profesional de la Información*, *16* (6), 254-262.

Sowa, J. F. (2000). *Knowledge representation: Logical, Philosophical and Computational Foundations*. Pacific Grove: Brooks/Cole Thompson Learning.

Stallman, R. (2004). *Software libre para una sociedad libre.* Madrid: Traficantes de Sueños.

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge Engineering: Principles and Methods. *Data & Knowledge Engineering*, *25*, 161-197.

Temmermman, R. (1999). Sociocognitive terminology theory, in: Feliú, J. & Cabré, M. T., *Terminología y Cognición: II Simposio internacional de verano de Terminología*, pp. 75-92. Barcelona: IULA-UPF.

Tramullas, J. (2005). Herramientas de software libre para la gestión. Hipertext.net. http://www.hipertext.net/web/pag258.htm [Consulted: 29/10/2008].

UNE (1990). Norma UNE-50-106-90. Directrices para el establecimiento y desarrollo de tesauros monolingües: *equivalent to* ISO 2788-1986. Madrid: AENOR.

Wildgen, W. (1994). *Process, Image and Meaning. A Realistic Model of the Meaning of Sentences and Narrative Texts*. Amsterdam: Benjamins.

Wildgen, W. (2008). From Lullus to Cognitive Semantics: The Evolution of a Theory of Semantic Fields. *University of Bremen*. http://www.bu.edu/wcp/Papers/Cogn/CognWild.htm [Consulted: 26/10/2008]

Zeng, M. (2005). *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*. Maryland.

About the Authors

Jose Antonio Moreiro

Jose Antonio Moreiro joined the Carlos III University of Madrid in 1991. is Professor at the Department of Library and Information Science of the University. Jose Antonio Moreiro is leader of the Information Engineering Group, where research is being applied on knowledge organization systems. He teaches Knowledge Organization Systems and Indexing techniques. Is author of 8 monographs and collaborated in other 12, thus like of 64 articles in international and national magazines. Has participated in two European projects and directed or collaborated in 8 nationals, also in 12 publication committees. Has 8 long stays in Latin American universities, and explained courses and seminaries in 28 European and Latin American universities, and in 20 professional institutions. It has oriented 20 doctoral theses in the domain of Library and Information Science.

Jorge Morato

He is currently a professor of Information Science in the Department of Computer Science at the Carlos III University of Madrid (Spain). He obtained his PhD in Library Science from the Carlos III University in 1999 on the subject of Knowledge Information Systems and its relationships with linguistics. Professor Morato has taught courses on Information Retrieval, Search Engine Optimization, Software Engineering, and Knowledge Modelling Techniques and Management Systems. From 1991-1999, he had grants or contracts from the Spanish National Research Council. His current research activity is centred on text mining, information extraction and pattern recognition, NLP, information retrieval, Web positioning, and Knowledge Organization Systems. He has published mainly on semi-automatic construction of thesauri and ontologies, topic maps, and conceptual and contextualized retrieval of semantic documents.

Sonia Sanchez-Cuadrado

In 2001, she received a research fellowship (Personal Research) from Spanish Government (MCYT), to a research project of Department of Information Science and Department of Informatics of Carlos III University of Madrid. From 2001 to 2003, she researched about Knowledge Organization Systems (KOS) and automatic construction and she received her PhD, with a work on Methontology to automatic construction of knowledge organization systems and Natural Language Processing (NLP). From 2003, she worked as Assistant Professor in the Department of Informatics at the Carlos III University of Madrid. Her main research subjects within Knowledge Reuse Group are a domain analysis and automatic construction of knowledge organization systems such as thesauri and ontologies. This works have been realized into a project followed by public institutions or private companies. Her current research activity is centred on text mining, information extraction and pattern recognition, NLP, information retrieval, Web positioning, and Knowledge Organization Systems. She has published mainly on semi-automatic construction of thesauri and ontologies, topic maps, and conceptual and contextualized retrieval of semantic documents.

Anabel Fraga

She is a Computer Engineering professional. Previous to set aside in the academic work, she committed her efforts in the industry as UNIX Administrator (HP-UX, Digital Unix, and so forth), Application Administrator (SICAP, Comptel technology for Telecom companies), Windows Administrator and Project Management. She obtained the E-commerce Msc. in the Carlos III of Madrid University. She is studying a PhD in Computer Science in the Carlos III of Madrid University. Her central areas of research are: Software Architecture, Information Engineering and Reuse; but she is also interested in ethics and innovative methods of learning supporting new software architects. She is currently professor of Software Engineering and Information Engineering in Carlos III of Madrid University. She is member of ACM CSTA and IASA.