THE EVOLUTION OF THESAURI AND THE HISTORY OF KNOWLEDGE ORGANIZATION: BETWEEN THE SWORD OF MAPPING KNOWLEDGE AND THE WALL OF KEEPING IT SIMPLE

Francisco-Javier García-Marco

Francisco-Javier García-Marco, Universidad de Zaragoza, Departamento de Ciencias de la Documentación e Historia de la Ciencia, Facultad de Filosofía y Letras, C/ Pedro Cerbuna 12, 50009 Zaragoza, jgarcia@unizar.es

Abstract

Thesauri are considered as an optimum between maximum ontological modelling (best knowledge mapping) and minimal alphabetic ordering (less expensive access). From this point of view a swift history of its evolution is provided. The recent evolutions in Internet searching are also analysed from this perspective. In this context, there is an immediate role for thesauri to ensure interoperability and feed up the new Internet semantic engines; and in the long term as a simple semantic user interface for resource discovery and navigation, which ensures proper transparency and control by the user who wants to take the effort to supervise and analyse its search processes. It is proposed that better devices for ensuring semantic sorting are provided when necessary, and that a distributed hub for thesauri interconnection is provided, perhaps using the existent big open Internet semantic facilities, as Wikipedia.

Keywords: Thesauri; Evolution; Interoperability; User experience; Cost-effectiveness; Semantic web

1. Introduction

During the last century and a half, many knowledge organization systems (KOS) —both models and exemplars— have been proposed and developed, but their success has been diverse, with some universal and specialized classifications and the thesaurus model being the main survivors. The context, history and current digital challenges of thesauri are well-known (Foskett, 1982; Dahlberg, 1995; Gilchrist, 2003, Dextre Clarke, 2008; Dextre Clarke & Zeng, 2012; Roe & Thomas, 2004) together with the most recent achievements, for example the new ISO standard (Dextre Clarke & Zeng, 2012), so this paper does not intend to review this evolution, but tries to re-examine some principles behind the wider historical frame of knowledge organization, so that their strengths can be better highlighted. The departure point of such an argument is the consideration of classification as the primary device that cognitive systems use to overcome the information overload, and to represent it in a way that is reusable, retrievable (García-Marco & Esteban-Navarro, 1993). In contrast to classification, which is natural, alphabetical ordering is a socio-technology, which appeared very late in the evolution of humanity, and which is valid for the manipulation of the keys when you know them, but cannot be used to map knowledge, but only to ensure a quicker access to it under certain conditions. After the invention of the alphabet, newer and newer technologies have appeared that are based on alphabetical order (alphabetical indexing), but these very efficient technologies can be considered only as a secondary approach to that of classification, one of the pillars on which human cognition is grounded.

But classification is hard work that increases when the knowledge that must be managed increases, so alternatives are needed when this work becomes overwhelming. So, the history of knowledge organization can be considered as a fluctuation between the (upper) ideal horizon of a complete ontology of knowledge and the (bottom) very practical and cost-efficient approaches based on accessing the alphabetically codified and ordered knowledge keys, disregarding the desire to properly map knowledge as something unrealistic and

even naïve. A middle point between both approaches could provide some stabilization to the knowledge organization trends, and this middle point could be precisely the thesaurus.

In this way of reasoning, we think that the strengths of thesauri, that have allowed them—at least up to now—to survive and thrive in the last decades, could very well be valid in the new Internet dominated global information space. These strengths can be summarized as the balancing of their representational power, on the one hand, and their simplicity on the other. To these, which are classical, we should add a new one of special significance in the age of the pervasive Internet search engines, which had been applied by almost all KOS up to the present: that is, transparency to both designers and cataloguers and, even more important in these times of apprehensions regarding digital privacy, users.

To stress this point, we approach to the history of thesauri as a mirror of the general evolution of knowledge organization, and as a function of its instrumental role for human memory, which needs keys and organization to preserve knowledge and access to it. Secondly, it studies the development of thesauri as an answer to the problems posed by uniterms and descriptors—the successively corrected post-coordinate indexing systems that were proposed as an alternative to enumerative classifications in a context of information explosion—; its merger with the alternative approach of synthetic classifications; and its great impact on other KOS. Finally, we try to contextualize and explore their role in the Internet 'world', the new knowledge infrastructure.

Thus, this approach is not empirical, but rationalist in the sense that it uses deductive reasoning to the field of KO by applying more general well-grounded theories from other fields that deal with the basic phenomena that KO studies, e. g., information, knowledge, language, etc. In particular we will try to root thesauri in the long-term view provided by the evolutionary history of nature and culture. The former (the rationalism-based perspective) is consistent with the majority of research trends in the field, as shown by Hiørland (2014), and, with him, we consider that it has been, and can keep being, a fruitful approach for producing sound working theories in our field; though, at the same time, we explicitly recognize the need for increased empirical research.

2. The primitive nature of classification

Though knowledge organization as a discipline has a specific sense related to bibliographical classification and other alternative approaches, knowledge organization is, more broadly, a natural process (García-Marco, 2011; García-Marco & Esteban, 1993). Organisms store representations of external and internal processes, so that they can operate on them, and obtain a result that can trigger a response that may have a relevant impact on their external or internal environment. These links between real-world configurations, internal representations and actions-in-the-world are governed by the laws of perception and memory. In particular. These links are organized in our memory in chains that are marked by keys, so that they can be accessed, distinguishing some from the others. But as human perception, human information processing and retrieval are very limited by factors as limitations of the processing space, available energy or computational complexity, so our systems are equipped to naturally discard knowledge that is not used, keeping the whole amount of operational knowledge to a certain and manageable size.

On the other hand, having as much knowledge available as possible gives a competitive advantage, because behaviour can be modulated in a more precise way. So humans have traditionally tried ways to expand their ability to remember and recall, applying their "crafting" abilities to this problem. Cognitive and educational psychologists commonly classify them in the realm of metacognitive abilities. For this discussion, it is important to note that, although they are biologically grounded, they are human inventions, assumed and institutionalized inside a culture.

The initial and most simple strategies must have been to select memory keys, able to be expanded in a "chunking" of others classified concepts (Miller, 1956). Humans learned very soon to map these mental keys against salient physical aspects of their environment (signals); and finally started to produce their own physical keys (symbols) and to manipulate their symbolic environments to create increasingly complex and expressive sets of signals or "discourses". They learned to link these institutionalized memories with the different astronomical events that formed repetitive cycles. Probably this increasingly complex relation of humans with their environment evolved together with the marvellous technology of language. (These processes have only been approached in the past by logic, in a speculative way, but they are now the object of empirical study by disciplines like Cognitive Archaeology (Wynn, 2002).) Language became itself a new field of experimentation and crafting. Rhythms, recitation and narration evolved as structures to support shared knowledge representation and recall. Representational techniques were slowly but impressively improved, as the history of fine arts shows; up to a point when the graphic world and language come together with the invention of ideographic script.

With the invention of writing, the process accelerated. Humans dealt not only with documents as a storage device, but also as a flexible working memory expansion, allowing them to improve notation and reasoning techniques. Logic and mathematics, libraries, information retrieval systems, the World Wide Web or Internet search engines are subsequent highly relevant

examples of these memory-enhancing techniques, roughly ordered in the line of historical evolution.

The aspect of this incredible evolution we want to stress at this point of the paper is that classification deserves careful examination as the first 'conceptual' technology that was invented to improve natural memory recall. (Other are emotional, as linking the representation with a feeling or emotion; or perceptual, as linking it with a more informative sign.) As Miller (1956) demonstrated in his article *The magical number seven, plus or minus two*, to overcome the capacity of the short time memory one of the best strategies is to use or invent concepts to create chunks of memory that can be accessed and processed together. Also Rosch (e.g. Lloyd & Rosch, 1978; Rosch, 1983) showed us that there are natural concepts, closely linked with natural perception, and that - over and bellow them - supra-ordinate and subordinate concepts are acquired later to manage and specify the previous natural ones to achieve better results in our actions. Later, formal techniques of classification and definition were developed, which are topics of teaching in disciplines such as logic or mathematics, and are intensely used in KOS.

Along our history, humans have become masters in these technologies of 'classification', which is at the foundations of philosophy (ontologies), sciences (taxonomies) and even at the roots of how our languages have evolved and grown. To master a field you must master its ontology, and be able to locate, though roughly, where a topic should be and which are its relevant relations to others. Education, the art of replicating our social shared knowledge for the new generations, is all about building these common conceptual frames, which will then be shared. This process of sharing the conceptual frames of a domain remains central for any community of practice (Albrechtsen & Jacob, 1998). In his brief and classical book, D. W. Langridge (1992) summarized it simply and powerfully: "There is no substitute for classification".

3. The two-dimensional nature of knowledge organization

Though rooted in classification as a natural process enhanced by a set of techniques, signs have at least two dimensions: one dealing with differentiating their form from other competing signs, so they can be distinguished properly (the 'signifier' dimension); and the other related to the mental or physical processes that such signs may elicit when interpreted (the 'signified´ dimension). In their conception of sign, Saussure (1916) and Peirce (1931) can be considered complementary, as the former insisted on the analysis of its differentiating internal structure, and the later stressed its pragmatic dimension, that is, its ultimate aim of supporting action.

Primarily, knowledgeable agents – such as human beings - are interested in meanings. So, as we have seen in the previous section, the first dimension over which humans have acted has been in organizing knowledge by conceptual order, creating more general and specific concepts and other specific relations among them. But with the beginning of external representation, humans discovered that they were able also to manipulate the perceptual properties of signs to order them. The most significant invention in this trend was the alphabet, which allows every verbal sign to be codified with very limited sub-signs (which is very efficient and connected to the way in which faceted classifications reduce the number of classes in the schedules). But the alphabet does something more, it brings also the possibility of ordering signs not by their meaning but by their form, e. g., the alphabetical order. In cognitive terms, this brought in a very cheap way of ordering information to access knowledge, because it requires no classification, no conceptual work. It is socially also very practical, because the community need not share totally the same conceptual hierarchy to retrieve or exchange, and one can delegate the manipulation of physical objects to other persons. With such an invention, big libraries and ultimately databases, information retrieval systems and search engines were made possible.

But anyway, though alphabetical ordering allows information retrieval, it does not produce the map that humans need to master a domain, to track it, move from one of its parts to another and, hopefully, to master it all.

So any good retrieval system must combine both of them: if you know the word or code that codifies a concept, alphanumerical systems are very efficient; if you do not know it exactly, you need a conceptual system of access. This is true for any knowledge store of some size (that we cannot memorize by heart), independently of its type. For example, good books have their tables of contents —the intellectual organization of the book, usually hierarchical or following any other system of organization— and the alphabetical index or indexes —the analytical ones. A traditional library has also the systematic arrangement and the alphabetical indexes. A web, at least after hypertext as a non-hierarchical system to relate documents proved a partial myth, has the search engine (alphabetical, but delegated) and the site map (conceptual and usually hierarchical, at least partially).

In conclusion, it seems that true knowledge organization is always conceptual; alphabetical access is only a secondary, though very useful, device. So, there is no reason to think that this will change in the foreseeable future, especially having strong cognitive scientific evidence to support it, both in the psychological and the cultural realms. Recently, Ingetraut Dahlberg (2011), in her *Ten Desiderata for Knowledge Organi-*

zation stated as the first and preliminary condition the "recognition of the fundamental difference in the organization of knowledge between the concept (i.e., the unit of knowledge)-the conceptual level-and the word, term or code-the verbal level-and the need for implementing this distinction in theory and practice (Desideratum 1)." And, specifically in the field of thesauri, Alan Gilchrist has insisted in different ways that one of the big advantages of thesauri is that it combines alphabetical and systematic access to information (e. g. Gilchrist & Kibby, 2000), an idea that has been stressed also by Guimarães (2003).

4. The dynamics of knowledge and its organization

Knowledge advances by increasing the network of concepts by discrimination, and also inventing wider concepts for manipulating other concepts that, up to a moment, were unconnected in such a way. They also grow in their inter-relations, be it at an analytical level (internal properties, dimensions or features) or synthetic (molar, existing or new aggregations). Besides cultural history and history of science, this has been also well-established in several fields of what is now cognitive science, such as psychology of learning, psychology of thinking, psycholinguistics and semantics. And in our field of information science, S. R. Ranganathan (1937) cleverly marked it as a keystone of his theory.

Knowledge must be organized, because otherwise it cannot be retrieved and reused. As we have seen, the psychology of memory has proved that at a psychological level, but when putting together several persons (and nowadays machines) to share knowledge, the same laws apply. But organizing is a work that grows in complexity and costs in direct proportion to the size and granularity of the domain to be ordered. To deal with this complexity, humans have invented different intellectual, organizational and instrumental technologies. These technologies have allowed human communities to access at new levels of complexity in their dealing with knowledge, only to see how, after some centuries or years, the new equilibrium collapsed under the weight of the new achievements.

As a result of these processes, the history of knowledge creation and organization has been cyclic. An age of rapid and dispersing discoveries and publications, like classic Greece, the Middle Ages expansion or the Renaissance, has been followed by another of consolidation and synthesis, like the Hellenistic period (Aristotelian synthesis), the High Middle Ages (Thomas Aquinas' *Summa*) or the Enlightenment (Diderot's French Encyclopaedia). This happened twice again in the last century, with the scientific upheaval of the world-wide wars and with the World Wide Web explosion.

In our modern times, the history of the *Encyclopædia Britannica* (Glasgow, 2002; Auchter, 1999, Encyclopaedia Britannica, 2014) exemplifies these tensions between knowledge creation and knowledge organization extremely well: from its first edition of two volumes in 1768-71, it grew very quickly in size to the 21 volumes and the first index volume in 1830–1842, to the 28 volumes of the 11th edition in 1910-1911. The physical size of the encyclopaedia reached a limit —an infrastructural one, related to the costs of distribution and storage— but the expansion of knowledge during the 20th century had to be addressed by reducing the wording of entries and considerably increasing them, up to 700.00 in the printed version of 2007. This huge alphabetically ordered repository of knowledge required a classificatory device to facilitate the navigation by the users, and also to produce a more compact less redundant content. Thus, the Propaedia outline of knowledge was introduced in the 15th edition of 1974-1984, notably after the great knowledge explosion of the world and cold wars. Finally, as a result of the digital information explosion another way of compiling and organizing knowledge for reference was needed, both more compact and usable and, thanks to the hypertext model (Bush, 1945), Wikipedia took a torch from the Britannica and other traditional encyclopaedias,

So, Britannica exemplifies how both the technologies that facilitate the increase in knowledge production and those that improve knowledge mapping and integration have been needed to maintain the growth of cultures.

As Jaenecke (1997) noted, the best knowledge organization is theory formation, as in this way all of the relevant concepts in an area of interest are defined and related in the most specific way; but a systematic theory-like arrangement of all the concepts that a culture manages is, at least for now, impossible (let us watch developments in the frame of the semantic web project). So knowledge organization is precisely the pursuit of a systematic arrangement of knowledge in an effective and efficient way, something that is at the same time impossible in its totality and absolutely necessary for retrieval, because it is the only way to allow the prediction of the position of a piece or record of knowledge inside a hypothetical knowledge base. A single and unique theory could be considered an optimum, but this can be maintained only for a period of time (Kuhn, 1970, scientific revolutions).

The use of what are now called ontologies reveals also this movement, as an ontology is a complete representation of a domain. But its use in the plural suggests the impossibility of achieving complete success over a wide domain (and certainly not a universe of knowledge in the sense of, for example, the UDC), so that competing ontologies must be considered and managed.

Retrieval in a big system of knowledge like Aristotle's synthesis or modern databases is all about ordering and

being able to predict the position of an item in the ordering system. Memory functions in the same way: once you have outstripped the capacity of the short-term memory, you need to classify to regain control, adding a label (a key) by which you may recall each class. Of course mapping knowledge in this way requires opting for a particular order when several are possible, and making the necessary references between related concepts that are separated in the schedule.

But in this age of knowledge expansion, it becomes very difficult to keep up with the organization of knowledge. For many years the only way to organize records was physically and visually in a space or temporarily on the calendar. After the invention of phonetic scripts, it was also possible to sort by the alphabet order; that is, by the 'signifier' of the sign. It became possible to sort by the first words of a record, by an abstract (a title may be nothing more than an abstract), or by other relevant properties, such as the author. But when the files grow too numerous, the individual and social cognitive systems cannot process them well, and begin to select and re-organize. New theoretical syntheses are produced, some of them become selected as references and the old authors are forgotten; subjects are mapped and new handbooks and encyclopaedias appear.

This cyclic nature of the dynamics of knowledge is consistent with those of other fields studied by very different sciences, which can be of an enormously different range of time-frames: earth climate fluctuations, sun spots, the business cycle, circadian rhythms, the pulsations of the heart, electromagnetic waves, light waves and so on. They have been related to the fluctuations of energy when it moves through an environment. In the case of knowledge organization, the source of energy would be the motivation to achieve a perfect organization of existing knowledge, a universal ontology comprising all concepts and their relations. The environment in which such an energy would degrade (and which it would invigorate) would be the particular societies in which knowledge organization is pursued, with their possibilities and challenges, and specifically the size and rate of increase of socially shared knowledge and the technologies available to cope with them. As we stated in the introduction of this paper, the emphasis in KOS development will fluctuate between the top line of achievable knowledge organization and the bottom one of applying the marvellous invention of alphabetical ordering to access at least the keys to access relevant pieces of knowledge.

5. Birth, growth and evolution of thesauri: reuniting the advantages of the alphabetical and hierarchical indexing

The evolution of thesauri (Foskett, 1982; Dahlberg, 1995; Gilchrist, 2003, Dextre Clarke, 2008; Roe & Thomas, 2004; Dextre Clarke & Zeng, 2012) follows also the lines that have been drawn in the previous chapter. They were born as systems to overcome the diversity of keywords and uniterms, a system invented to try to reign over the proliferation of scientific research and literature in the age of the world wars (that made bibliographic classifications seem obsolete), using the metaphor of the dictionary (an alphabetical system).

After simple aggregated lexical tags proved unsuitable for many purposes, classification by disciplines was reintroduced in thesauri (note the apparent step back in the direction of universal classifications). After that or more precisely during that effort, the explosion of interdisciplinary research put disciplines into question, and universal facets triumphed as a system to organize the new complexity. Thesauri proved also useful in reconciling subject-headings (based on alphabetical indexing) with classifications, able to switch between them better; something not so strange, as thesauri had walked this very way before.

As the knowledge explosion grew – not only in the scientific research area, but also in the media, managerial activities and the arts, other systems were invented to deal with the proliferation of knowledge records that competed with thesauri, because they were also able to add to the conceptual dimension, though in an indirect, automating and deterministic way, mainly using probabilistic search according to co-occurrence and "social search" (citation or reference). But still they do not provide the conceptual maps that one needs to search by the more important dimension of knowledge, which is the conceptual one.

Meanwhile, other more complex systems (including some complex thesauri structures) were abandoned, such as the sophisticated and beautiful PRECIS (1), because the work they required when indexing and retrieving were only suitable for very specific fields, in which indexing must incorporate many dimensions (e. g., Chemistry or Genetics), and more importantly, the added value to the information processes paid for it.

After their initial impressive growth, thesauri had a big impact in other KOS with an orientation towards vocabulary control and alphabetical access. Very important subject headings lists and other authority lists were recast using thesaurus relations and approaches. LCSH was a pioneer in the eighties (Stone, 2000), and now it is available in SKOS and other semantic web compliant formats, as are the other LC vocabularies (Library of Congress, 2015a, 2015b). Following a similar history as thesauri but in an accelerated way, facet analysis was finally also incorporated in the effort to make LC authority tools more compact and usable (project FAST, Chan & O'Neill, 2010).

In a different arena, thesauri and classifications have been consistently approaching each other during all

these years (see remarkably Thesaurofacet, Aitchison, 1970), though not in the same degree as subject headings and thesauri, which share their basic alphabetical approach. In thesauri, hierarchical relations were reintroduced as compulsory to interconnect the concepts; and classificatory approaches to the arrangement of the vocabulary were also codified in the standards (disciplinary, domain based…). On the other hand, classifications have adopted the controlled coordination of facets as a way to reduce the number of actual classes in the schedules and deal with the explosion of subjects, though tensions have arisen around the need to preserve a unique way of arranging documents (and to ensure user prediction of a class position), which only enumerative classifications or faceted ones with strict synthesis rules can achieve. Facet analysis (Ranganathan, 1937) has clearly been a key factor in this mutual endeavour, and resulted in early efforts to achieve a synthesis between thesauri and classifications, as the Thesaurofacet (Aitchison, 1970) and classauri (Devadasson, 1985). As Dahlberg has repeatedly insisted on, knowledge organization systems must be able to express the concepts to be retrieved, which are always evolving, and this requires a faceted system that isolates the single concepts that will therefore be combined in precoordination or for postcoordination (1995).

Though the potential of thesauri as the tool to achieve the much desired "unified theory of knowledge organization" was seen very early on, the new ISO standard has set clearly and practically the ground to establish thesauri as a conceptual and standardized spine for the rest of the KOS, which is very much needed in the new Internet environment and which, by its very nature, does not accept local or idiosyncratic procedures for universal tools. For this purpose, rules for deriving consistent conceptual arrangements from thesauri — allowing classifications to be expressed as thesauri, or if needed, for thesauri to be used as classifications must be carefully standardized, which could be a challenge for the next edition of ISO 25964.

6. Thesauri and knowledge organization in the age of Google

In our modest opinion, the history of Internet search, though shorter, resembles an accelerated recapitulation of the previous century of knowledge organization efforts.

In the first years of the Internet, conceptual access (hierarchical) was the rule: we can remember FTP services, bulletin boards or Gopher services, strictly hierarchical. For some years, even the most popular service for resource discovery in the World Wide Web was a classificatory one, that is, Yahoo categories. The explosion in the number of web pages broke the back of these systems, which relied on human work, and a combination of keyword extraction and relevance weight triumphed. Probabilistic search uses mainly two approaches to calculate relevance: information, that is, words that are rarer in the collection; and also redundancy, computing the number of the same words in a text. But, when the collection grows, the results typically become correspondingly poorer; so the next crisis appeared. Google overcame the problems of the model adding to it social indexing—the number of incoming links—, the computed form of the classical device of intertextual citation, cleverly used by Garfield for impact indexes. With PageRank, Google's approach clearly triumphed over their competitors and now has a global market share of 66,74%, with Bing, Yahoo and Baidu following behind with around 10% each, according to Net Applications (2015); and 89,62% according to the statistics aggregator Statista.com (2015). Even in academic contexts, simple keyword Google searches seem to be increasingly preferred by students (Georgas, 2015) and even teachers (e. g. Kemman, Kleppe & Scagliola, 2012).

But it is not only that users prefer Google when searching the Internet; they also use it in the simplest possible way. Most of the research shows that a great majority of searchers do not go beyond those keywords that naturally come to their minds, do not examine the concepts carefully to find synonyms or related terms, do not use commands, do not expand or refine their searches, do not examine metadata and do not look for other results than those which are in the first page. In this context, it is a legitimate question to ask if metadata, thesauri, and in general, information control devices deserve the effort and costs that they require.

Without further research, we must remember first of all that, as in many human activities, the cost of information control will only be accepted if the expected benefits exceed the cost, or when the costs of misinformation are expected to result in a greater pain; and this felt only by those persons who are sensible to the benefits and risks, and are prepared and willing to make the effort. In many cases, such people are in charge, and few will consider making the effort to disambiguate persons with the same first and second name in the database of a hospital, jail or police station; or businesses with the same name in a tax-oriented database. On the contrary, critical databases such as those of chemical or pharmaceutical products are heavily controlled, and their vocabularies carefully connected. So information control naturally thrives in those contexts where information has a critical value because of economic, political or cultural reasons.

In other contexts, information is not critical but only useful, and it is so redundant that it can be easily obtained by the previously described automated methods. However, it must be remembered here that, though most of the users do no want to make the effort to get better answers, they can usually distinguish very well

between two different search sets how relevant they are to their needs, and are able to say which one served their interests better. Because users notice well the differences in quality, search engines providers are bond to compete. So they have a competitive pressure to build in the kind of features that improve search results in the ever-expanding (at least for the moment) galaxy of digital information. When users are given an easy tool that provides better results for an identifiable activity, they naturally adopt it, as Helen Georgas (2013, 2014) showed when comparing the preferences of undergraduate students between federated search and Google.

In fact, Google has been rapidly incorporating an explicit semantic dimension to its search model. In February 2004, Google improved its Latent Semantic Indexing (LSI) model expanding its ability to understand synonyms. (Mooz, 2015). In May 2012 Google began the implementation of its "Knowledge Graph" that codifies people, places and things, and the relations among them, that was remarkably presented as retrieving "things, not strings" (Google, 2012). So, before 2012 concepts and their relations as something more than the strings that act as "signifiers" had actually made a notable entrance in the strategic plans of the leader search engine. Also around these dates, the main search engine operators were busy trying to establish some common ground for cataloguing resources with an economic value, able to pay for the effort (mainly shopping, travelling, etc.), and the schema.org (2012) initiative was launched. So, users were effectively asking for more, and the Internet information providers were ready to try to comply, starting of course with the sectors in which their efforts would be more probably repaid. But a decade before, the leaders of the World Wide Web had actually stated very clearly that the Internet would be broken by its own growth and weight if not given more structure, proposing a multi-year project, which was meaningfully called the semantic web (Berners-Lee, Hendler & Lassila, 2001).

So, the Internet is becoming semantic, recognizing that the "significant" exists and that string-oriented searches do not satisfy the users if they do not incorporate concepts as intermediate entities. However, such a moral victory could very well be the tomb of thesauri and other semantic-based indexing and search tools. If Google and other search engines incorporate complex ontologies to enhance searches and user experience, what role could be left for thesauri? Two different ones could be envisaged.

The first one is clearly transitional: to provide a reliable interoperable standard to feed the knowledge bases of the search engines from the existing knowledge repositories. Through the clear formalization that ISO 25496 provides, concepts and relations can be easily and automatically imported into the search engine systems.

But there is a more fundamental possibility: empowering the knowledge user. Search comfort and efficiency are certainly great values, but they are not the only ones to be considered in the long term, if we want to preserve good and universal access to knowledge. Transparency is also a key factor in this direction, and one of the most relevant aspects of knowledge maps versus search algorithms is precisely that transparency. Knowledge maps are there for everybody to discuss; but algorithms are increasingly proprietary and secret, only tips are provided about them to the public. In a previously cited work, Kemman, Kleppe & Scagliola (2012) raised this topic, indicating that scholars are becoming increasingly dependent on "black boxed algorithms", calling into question the academic principles of provenance and context.

Thesauri may be especially useful for solving the transparency gap by providing concepts, relations and terms in clear presentations, easily to be understood by users. For this purpose, the balance between representational power and simplicity could be decisive. The thesaurus model is powerful and easy to explain, providing a perfect tool to increase transparency and to empower the user or, at least, those who need and want more control over the information they use; certainly the minority that can become more vocal in defending their rights and perhaps in the process embarrass the largest Internet players. Certainly, thesauri and KOS in the Internet should include devices to codify and store these discussions about knowledge maps, and this is a challenge that will probably be addressed in the new editions of the ISO 25964 standard.

7. The need for thesaurus 'hubs' and the role of social technologies

Though the strategy to interconnect thesauri using the semantic web that is explicit in the ISO 25964 is a great bottom-to-top device, it would be advisable to try to complement it with a top-to-bottom strategy. Interconnecting the exiting KOS is a great approach to improve retrieval across the Internet, especially all those well-catalogued resources that the open data movement and the semantic web are contributing to, but there is a need for a hub where concepts and their relations can be stored in a transparent and friendly way.

Dreaming of a single KOS is just a dream, but it can be a very useful one, as the ecology of knowledge organization needs "animals" of different sizes and kinds to thrive. Big search engines are doubtless such big animals, but they are black boxes for deeper technical work, at least up to the moment.

In a previous paper (García Marco, 2016) we explored in depth the possibility that an extended Wikipedia could constitute such a platform:

> After all, conceptual dictionaries, encyclopaedias and thesauri are very close siblings in their etymology and history. Strikingly similar approaches between the two worlds are relatively recent, as the history of Propaedia and the invention of "conceptual" thesauri (thesauri including definitions) show. The underlying etymology of 'treasury' (from the Greek θησαυρός, thesauros, storehouse, treasure) that both worlds share denotes an effort to select the best of both worlds—a language, a terminology, a collection of concepts, citations or texts— and offer it in an organized way, usually not only alphabetical but systematic and with devices for controlling synonymy and polysemy. Only their immediate aims differ: encyclopaedias facilitate learning and education; dictionaries, writing and reading; and thesauri, retrieving from repositories, catalogues and documents.

In fact, knowledge experts from different semantic-related and many other disciplines are actually exploiting the huge semantic repository that Wikipedia has become (Okoli et al., 2014; Mesgari et al., 2015). Of course, part of this research has been related to enhancing document clustering and information retrieval (e.g., Hu et al., 2009). Also, some scientific communities have been exploring the potentialities of the Wikipedia as a social semantic hub, notably in the field of genome research (e.g., Gardner et al., 2011).

Such a strategy –or a similar one- would have the problem of the difficult and slow institutional work that would be necessary. But the legal frame is available, and the interconnected thesauri would not need to reside inside the actual wiki, complicating Wikipedia management. They can be connected through semantic web standards, providing they have a permanent updating engine and persistent addresses. Of course, recognition, copyright and copyleft compliance should also be ensured, and some kind of a more formal cooperation would be convenient. The key aspect of an approach like this would be to anchor vocabularies to a big encyclopaedic scheme, used and supported worldwide. An international organization such as ISKO could promote and support some kind of 'World Wide Web Thesaurus Foundation' to bring together thesauri and other KOS editors and users, together with the World Wide Web Consortium and the larger KOS editors.

It must be also taken into account that, in fact, Wikipedia is more useful as a terminological and encyclopaedic database than as an explicit categorical system; because the ontologies are frequently idiosyncratic, not based on the existence of sufficient definitions or even confusing (Zazo et al., 2015). So, 'anchored' KOS could serve also as potential alternative systematic access tools to Wikipedia, reciprocally benefiting the Wikipedia project.

Of course, such an approach would require that the terms of a thesaurus that are not present in Wikipedia or the selected hub are given the corresponding entry in the wiki. So, updates and new terms in Wikipedia should be proposed and edited by the thesaurus managers, contributors and editors to preserve consistency, as new concepts will be found. It would be also a great way to contribute to the Wikipedia project, the most widely used reference tool in the world.

So, using a large social semantic medium like Wikipedia or similar as an anchor could serve both ends of the knowledge organization ecology. On the one hand, it would provide a wide vocabulary whereon to hang semantic web compliant thesauri and other SKOS, functioning as a big bi-directional hub. On the other hand, it would serve to take advantage of social knowledge sharing and editing, and become a source of data about this tagging and linking activity that would be very useful for evaluation purposes and for enhancing relevance judgements. Apart from the complementary dynamics between the top-down and bottom-up approaches, such a semantic network of KOS around a large tool such as Wikipedia would be very easy to exploit by the 'big animals' of internet search, ensuring that thesauri and other KOS do not become isolated in the emerging Web 3.0, and that they find their way forward. In the aforementioned paper (García Marco, 2016) a deeper discussion of the expected benefits and technical issues is provided.

8. Conclusion and proposals

To sum up, thesauri represent a good compromise between the quest for knowledge organization and the need to keep it simple, economic, efficient and dimensioned to the required specificity and relational density of the concept maps that must be used to master a field or to exploit it. The new thesauri standards are a big jump precisely in this direction, connecting them with other KOS, and expressing them in a way compatible with networked computer systems, especially relational databases and the Internet. Knowledge organization is a need, and thesauri are one of their best tools.

Of course, there is work to be done if the promise of thesauri to become the leader in the KOS race in the digital age is going to be completely fulfilled. Without losing its simplicity and clarity, the thesaurus standard must become increasingly hospitable to other KOS characteristics, notably the need to produce predictable strings and codifications to order things and documents hierarchically, be it in the physical realm —shelves, exhibitions…— or in the digital one —systematic and semi-systematic lists for browsing, mapping and discovery. They should also maintain compatibility with competing standards for graphically mapping and exploring resources, like topic maps. Knowledge visualization and graphic representation is becoming increasingly popular in the age of big data, and it will be a big challenge to any future thesaurus standard to incorporate these new presentation formats that are originating in other fields of practice, but which can improve the visual experience of the users.

Also, there is a need for a hub where thesauri and other KOS might be connected, searched and exploited in different ways. The semantic web allows for completely distributed interconnection, but humans and their deeds need reference points. This task should be pursued by the biggest KOS editors and in conjunction with the big Internet institutions and firms. Wikipedia, because of its size, popularity and free foundation status, could be a good candidate. KOS must be opened to social collaboration and social tagging, but there is also a need for the accumulation and preservation of efforts. In this respect, Wikipedia has also proved to be a leader.

Complex ontologies are probably the highest stage of knowledge organization that can now be seen–as formal representations of strong theories about a domain, in the sense of Jaenecke (1997)-, but thesauri are for the moment a simpler and cheaper way to start the semantic journey in most fields and activities; and probably, even in the future, they will still be a simple and practical model for information systems to communicate with normal users even if a strong ontology lies behind. Even if thesauri are overtaken by complex ontologies in those fields in which they may be profitable, and because of their power, clarity and simplicity of their model, thesauri can still be the *de facto* standard to connect complex ontologies to the users, so they can pursue more transparent and sensible resource discovery in the increasingly complex digital world.

But though semantic technologies are central to the future of thesauri, knowledge engineers should not be the only focus of attention of the knowledge organization community. In particular the contacts with the lexicological and terminological communities that used to be traditional in our field (Dahlberg, 1991, 1992) should be reinforced. These communities are also busily involved in building or using lexical thesauri around Wikipedia to overcome very similar problems to those of us, and not only with a generic linguistic approach but also in very specific scientific domains (e.g., using Wikisaurus, Rapisardi & Giardino, 2014).

As for knowledge organization as a discipline, we have a clear future in the theoretical field as an organized store of knowledge on the evolution of the human efforts to keep knowledge represented, stored, organized and retrievable. Also in the practical field, so long as we keep learning and discussing with other partners in this quest, which is now being led by large organizations, some outside the Library and Information Science field such as Google, Apple, IBM, Microsoft, as well as others that are clearly inside, such as the Library of Congress or Europeana, among others. On the practical and applied side, we could be momentarily overwhelmed, if we do not master the new computing technologies swiftly enough. But even in the worst scenario, as these fields specialize again in their different subfields, the classic labels and disciplines will slowly emerge back, because the domains in which they research are not temporary, but a constant in the relation of humans to their environment.

Notes

(1) PRECIS was commissioned by the British National Bibliography (now part of the British Library) but was finally abandoned because in application it was too expensive (Austin & Butcher, 1969). It was intended as a universal classification scheme.

Acknowledgments

This research is part of the ongoing project CSO2015-65448-R, Possibilities and requirements of knowledge organization systems for interoperability between memory institutions and the sector of cultural tourism, funded by the Ministry of Economy and Competitiveness of the Government of Spain, R&D&I 2015 National Program of Research, Development and Innovation Oriented to the Challenges of Society.

I want to acknowledge especially the suggestions of Alan Gilchrist and Ingetraut Dahlberg who commented on the paper at different moments and the anonymous referees. They are not responsible however for any of the faults that may be found in this paper.

References

Aitchison, J. (1970). Thesaurofacet: a multipurpose retrieval language tool. // Journal of Documentation 26:3 (1970) 187-203. doi:10.1108/eb026493.

Albrechtsen, H.; Jacob, E. K. (1998). Classification systems as boundary objects in diverse information ecologies. Medford: Information Today Inc, 1998.

Auchter, Dorothy (1999). The evolution of the Encyclopaedia Britannica: from the Macropaedia to Britannica Online. // Reference Services Review 27:3 (1999) 291-299.

Austin, D. W.; Butcher, P. (1969). PRECIS: a rotated subject index system. London, British National Bibliography, 1969.

Berners-Lee, T.; Hendler, J.; Lassila, O. (2001). The semantic web. // Scientific American 284:5 (2001) 76-88.

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

Chan, L. M.; O'Neill, E. T. (2010). FAST: Faceted Application of Subject Terminology: principles and applications. Santa Barbara, Calif.: Libraries Unlimited, 2010.

Dahlberg, I. (1991). Knowledge organization, thesauri, and terminology. // International Classification 18:3 (1991) 133.

Dahlberg, I. (1992). Knowledge organization and terminology - philosophical and linguistic bases. // International Classification 19:2 (1992) 65-71.

Dahlberg, I. (1995). Current trends in knowledge organization Organización del conocimiento en sistemas de información y documentación: actas del I Encuentro de ISKO-España, Madrid, 4 y 5 de noviembre de 1993: Zaragoza : [ISKO-España : Universidad de Zaragoza], 1995, 7-26. http://www.iskoiberico.org/wp-content/uploads/2014/07/007-026_Dahlberg.pdf.

Dahlberg, I. (2011). How to Improve ISKO's Standing: Ten Desiderata for Knowledge Organization. // Knowledge Organization, 38:1 (2011) 68-74.

Devadason, F. J. (1985). Online construction of alphabetic classaurus: a vocabulary control and indexing tool. // Information Processing & Management 21:1 (1985) 11-26.

Dextre Clarke, S. (2008). The last 50 years of knowledge organization: a journey through my personal archives. // Journal of Information Science 34:4 (2008) 427-437. doi:10.1177/0165551508089225.

Dextre Clarke, S.; Zeng, M. L. (2012). From ISO 2788 to ISO 25964: the evolution of thesaurus standards towards interoperability and data modeling. // Information standards quarterly 24:1 (2012) 20-26.

Encyclopaedia Britannica. (2014) // World Heritage Encyclopedia. http://community.worldheritage.org/articles/Encyclopædia_Britannica.

Foskett, A. C. (1982). The subject approach to information (4th ed. ed.). London: Bingley, 1982.

García Marco, F. J.; Esteban Navarro, M. A. (1993). On some Contributions of the Cognitive Sciences and Epistemology to a Theory of Classification. // Knowledge Organization 20:3 (1993) 126-132.

García Marco, F.-J. (2011). The information pyramid revisited: enriching the cognitive sciences model. // Profesional de la Informacion, 20:1 (2011) 11-24. doi:10.3145/epi.2011.ene.02.

García Marco, F.-J. (2016). Enhancing the visibility and relevance of thesauri in the web: searching for a 'hub' in the linked data environment. // Knowledge Organization, accepted for publication for the Special Issue on Thesauri.

Gardner, PP; Daub, J; Tate, J; Moore, BL; Osuch, IH; Griffiths-Jones, S; Finn, RD; Nawrocki, EP; Kolbe, DL; Eddy, SR; Bateman, A. (2011). Rfam: Wikipedia, clans and the "decimal" release. // Nucleic Acids Res 39 (2011 Jan) D141-5. doi: 10.1093/nar/gkq1129.

Georgas, Helen (2013). Google vs. the Library: Student Preferences and Perceptions When Doing Research Using Google and a Federated Search Tool. // Portal: Libraries and the Academy 13:2 (April 2013) 165-185. http://muse.jhu.edu/journals/portal_libraries_and_the_academy/summary/v013/13.2.georgas.html.

Georgas, Helen (2014). Google vs. the Library (Part II): Student Search Patterns and Behaviors when Using Google and a Federated Search Tool. // Portal: Libraries and the Academy 14:4 (October 2014) 503-532.| http://muse.jhu.edu/journals/portal_libraries_and_the_academy/summary/v014/14.4.georgas.html.

Georgas, Helen (2015). Google vs. the Library (Part III): Assessing the Quality of Sources Found by Undergraduates. // Portal: Libraries and the Academy 15:1 (2015) 133-161. https://muse.jhu.edu/.

Gilchrist, A. (2003). Thesauri, taxonomies and ontologies - an etymological note. // Journal of Documentation 59:1 (2003) 7-18. doi:10.1108/00220410310457984.

Gilchrist, A., & Kibby, P. (2000). Taxonomies for business: access and connectivity in a wired world: TFPL Limited, 2000.

Glasgow, Eric (2002). Scotland and the Encyclopaedia Britannica. // Library Review 51:5 (2002) 263–267. http://dx.doi.org/10.1108/00242530210428764.

Google (2012). Introducing the Knowledge Graph: things, not strings. // Google Official Blog, May 16, 2012. https://googleblog.blogspot.com.es/2012/05/introducing-knowledge-graph-things-not.html.

Guimarães, J.A.C. (2003). A análise documentária no âmbito do tratamento da informação: elementos históricos e conceituais. // Rodrigues, J.M.; Lopes, I.L. (org.) Organização e representação do conhecimento na perspectiva da ciência da informação. Brasília: Thesaurus, 2003. (Estudos avançados em ciência da informação; 2). 100-117.

Hu, Xiaohua; Zhang, Xiaodan; Lu, Caimei; Park, E. K.; Zhou, Xiaohua (2009). Exploiting Wikipedia as External Knowledge for Document Clustering. KDD-09: 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, FRANCE, JUN 28-JUL 01, 2009. 389-396.

ISO 25964-1:2011. Information and documentation. Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval. Geneva: International Society for Standardization.

ISO 25964-2:2013. Information and documentation. Thesauri and interoperability with other vocabularies. Part 2: Interoperability with other vocabularies Geneva: International Society for Standardization.

Jaenecke, P. (1997). Knowledge organization due theory to theory formation. // Organización del Conocimiento en Sistemas de Información y Documentación 2 (1997) 39-55.

Kemman, Max; Kleppe, Martijn; Scagliola, Stef (2014). 'Just Google It'. // Mills, Clare; Pidd, Michael; Ward (eds.) Proceedings of the Digital Humanities Congress 2012. Studies in the Digital Humanities. Sheffield: HRI Online Publications, 2014. http://www.hrionline.ac.uk/openbook/chapter/dhc2012-kemman

Kuhn, T. S. (1970). The structure of scientific revolutions (2nd ed. ed.). Chicago: University of Chicago Press, 1970.

Langridge, D. W. (1991). Classification: its kinds, elements, systems and applications: Bowker-Saur, 1991.

Library of Congress (2015a). Library of Congress Subject Headings. http://id.loc.gov/authorities/subjects.html.

Library of Congress (2015b). LC Linked Data Service: Authorities and Vocabularies. http://id.loc.gov.

Lloyd, B. B.; Rosch, E.; Social Science Research Council (U.S.). (1978). Cognition and categorization. Hillsdale, N.J.: Lawrence Erlbaum, 1978.

Mesgari, M.; Okoli, C.; Mehdi, M.; Nielsen, F.A.; Lanamaki, A. (2015). The Sum of All Human Knowledge: A Systematic Review of Scholarly Research on the Content of Wikipedia. // Journal of the Association for Information Science and Technology 66:2 (2015) 219-245. doi:10.1002/asi.23172.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. // Psychological Review 63 (1956) 81-97.

Mooz (2015). Google Algorithm Change History. Learn Seo. https://moz.com/google-algorithm-change.

Net Applications.com (2015). Desktop Search Engine Market Share. // netmarketshare.com. Irvine, CA. Net Applications, August 2015. https://www.netmarketshare.com/search-engine-market-share.aspx?qprid=4&qpcustomd=0.

Okoli, C.; Mesgari, M.; Mehdi, M.; Nielsen, F. A.; Lanamaki, A. (2014). Wikipedia in the Eyes of Its Beholders: A Systematic Review of Scholarly Research on Wikipedia Readers and Readership. // Journal of the Association for Information Science and Technology 65:121 (2014) 2381-2403. doi:10.1002/asi.23162.

Peirce, C. S.; Hartshorne, C.; Weis, P. (1931). Collected papers of Charles Sanders Peirce. Cambridge (Massachusetts): Belknap Press of Harvard University Press. Vol. I-II, 1931.

Ranganathan, S. R. (1937). Prolegomena to library classification. Madras, London: The Madras library association; E. Goldston, ltd, 1937.

Rapisardi, E.; Di Franco, S.; Giardino, M. (2014). Web Participatory Framework for Disaster Resilience: Coping with Information Deluge. Engineering Geology for Society and Territory, Vol 7: Education, Professional Ethics and Public Recognition of Engineering Geology. New York: Springer, 2014.

Roe, Sandra K.; Thomas, Alan R. (2004). The Thesaurus: Review, Renaissance, and Revision. Haworth Press, New York, 2004.

Rosch, E. (1983). Prototype Classification and Logical Classification: The Two Systems. // Scholnick, E.K. (ed), New Trends in

Conceptual Representation: Challenges to Piaget's Theory?. Hillsdale: Lawrence Erlbaum Associates, 73-86

Saussure, F. d. (1916). Cours de linguistique générale. Paris: Payot, 1916.

Schema.org (2012). What is Schema.org? The type hierarchy. http://schema.org.

Statista (2015). Worldwide market share of leading search engines from January 2010 to July 2015. Statista: the Statistics Portal. New York: Statista Inc. http://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

Stone, A. T. (2000). The LCSH Century: A Brief History of the Library of Congress Subject Headings, and Introduction to the Centennial Essays. // Cataloging & Classification Quarterly 29:1-2 (2000) 1-15. doi:10.1300/J104v29n01_01

Wynn, T. (2002). Archaeology and cognitive evolution. // Behavioral and Brain Sciences 25:3 (2002) 389-402.

Zazo Rodríguez, A. F.; Figuerola, C. G.; Alonso Berrocal, J. L. (2015). Edición de contenidos en un entorno colaborativo: el caso de la Wikipedia en español. // Scire. 21:2 (jul.-dic. 2015) 57-67. http://www.ibersid.eu/ojs/index.php/scire/article/view/4243.

Received: 2016-02-01. Accepted: 2016-02-05