Documentary languages and knowledge organization systems in the context of the semantic web[1]

Linguagens documentárias e sistemas de organização do conhecimento no contexto da web semântica

Marilda Lopes Ginez de LARA[2]

Abstract

The aim of this study was to discuss the need for formal documentary languages as a condition for it to function in the Semantic Web. Based on a bibliographic review, Linked Open Data is presented as an initial condition for the operationalization of the Semantic Web, similar to the movement of Linked Open Vocabularies that aimed to promote interoperability among vocabularies. We highlight the Simple Knowledge Organization System format by analyzing its main characteristics and presenting the new standard ISO 25964-1/2:2011/2012 -Thesauri and interoperability with other vocabularies, that revises previous recommendations, adding requirements for the interoperability and mapping of vocabularies. We discuss conceptual problems in the formalization of vocabularies and the need to invest critically in its operationalization, suggesting alternatives to harness the mapping of vocabularies.

Keywords: Documentary languages. Interoperability. Knowledge organization systems. Semantic web. Thesauri. Vocabularies.

Resumo

Pretende-se discutir neste artigo a necessidade de formalização das linguagens documentárias como condição para seu funcionamento na Web *Semântica. Com base em revisão bibliográfica, apresenta-se a iniciativa do* Linked Open Data *como condição inicial para a operacionalização do projeto e, de forma associada, o movimento do* Linked Open Vocabularies*, voltado à promoção da interoperabilidade entre vocabulários. Destaca-se o formato* Simple Knowledge Organization Systems *analisando suas características principais e apresenta-se a nova norma* ISO 25964-1/2:2011/2012 - Thesauri and interoperability with other vocabularies*, que revê as anteriores acrescentando requisitos para a interoperabilidade e o mapeamento de vocabulários. Pontua-se problemas conceituais a serem enfrentados na formalização dos vocabulários e conclui-se pela necessidade de investir criticamente em sua operacionalização, sugerindo alternativas para aproveitar o mapeamento de vocabulários.*

Palavras-chave: *Linguagens documentárias. Interoperabilidade. Sistemas de organização do conhecimento.* Web *semântica. Tesauros. Vocabulários.*

Introduction

Knowledge Organization Systems, a term adopted by the International Society of Knowledge Organization (ISKO), consist of several tools that deal with the 'content' of documents, including documentary languages. Thus, the functions of the Semantic Web require specific formalization in to make them identifiable and interoperable. Initiatives related to the implementation of this formalization include open data, Linked Open Data

[1] Partial result of the research conducted in the senior internship at the Universidad Carlos III de Madrid under the supervision of Prof. Dr. José Antonio Moreiro González, with support from *Conselho Nacional de Desenvolvimento Científico e Tecnológico*, National Agency for Scientific and Technological Development, Protocol nº 201692/2011-2.

[2] Professora Doutora, Universidade de São Paulo, Escola de Comunicação e Artes, Departamento de Biblioteconomia e Documentação. Av. Prof. Lúcio Martins Rodrigues, 443, Cidade Universitária, 05508, São Paulo, SP, Brasil. *E-mail*: <larama@usp.br>.

Received on 10/4/2012 and approved for publication on 1/10/2013.

(LOD), the movement of the Linked Open Vocabularies (LOV), the proposal of the Simple Knowledge Organization Systems (SKOS) format and the recent documentary standard of the International Standard Organization (ISO) 25964-1/2:2011/2012, Thesauri and interoperability with other vocabularies.

Based on a bibliographic review, the aim of this article was to discuss the initiatives and critically examine their implementation taking into consideration the assumptions that underlie the Knowledge Organization Systems and open possibilities for its reuse within the Semantic Web.

Linked Open Data and Linked Open Vocabularies

Unlike the 'syntactic' Web - also called the hypertext Web -, that is based on the coincidence of characters, the aim of Semantic Web is to connect structured data (Berners-Lee, 2006). The first condition required for its development is open access to data, a step ahead of the Open Access initiative, whose aim was to promote access to scientific information, which is not less important.

Linked Data, or Linked Open Data, *Datos Abiertos Vinculados* in Spanish (Peset *et al*., 2011), refers to a set of procedures designed to promote open data to enable preparation, delivery and reuse. In the same spirit as Linked Open Vocabularies that seek to promote the publication of vocabulary-related open data to contemplate, more specifically, "A subset of a confusing LOD cloud" (Méndez; Greenberg, 2012, p.240) dedicated to the Knowledge Organization Systems (KOS).

The terms and principles of the Linked Open Vocabularies initiative are related to the family of Word Wide Web Consortium (W3C) recommendations for Semantic Web construction. Data is identified with the Resource Description Framework (RDF) standard, a resource description framework for metadata developed by W3C (Méndez, 1999), which includes an indication of the relation that may exist between this and other data, according to the set of triples consisting of subject, object, and predicate. It is possible to describe different types of data and subjects by coding each single piece of data with Uniform Resource Identifiers (URI).

Broader initiatives within Open Data include, among others, many details, vocabulary catalogues, such as the Data Hub, which are made available with specifications for reuse. Concrete experiences that have already been developed include the formalization of authority lists/records in the Library Congress, Library Congress Subject Headings, Thesaurus of the Food and Agriculture Organization of the United Nations (FAO) - known as AGROVOC -, and the list of *Encabezamientos Materia* - National Library of Spain (BNE) -, among others. The new KOS ecosystem not only allows for vocabulary to be put 'on the' Web for people to read, but 'into' the Semantic Web, allowing machines to use [the vocabularies] directly (Méndez; Greenberg, 2012).

Knowledge Organization Systems and Simple Knowledge Organization System

The term Knowledge Organization Systems (KOS) was initially used by Hodge in 2000 to encompass all types of information and knowledge management organizational planning, beginning with systems of classification, categorization, subject headings, authority lists, thesauri, semantic networks and ontologies (Hjorland, 2008). This was also used by Soergel (2001) who in a previous article had already established the need to use Uniform Resource Identifiers (URI) to foster an exchange of knowledge organization systems.

The Simple Knowledge Organization System (SKOS), in turn, is a recommendation of the W3C that was developed in the spirit of Linked Data (World Wide Web Consortium, 2012). It consists of a data model for sharing and linking knowledge organization systems via the Web (Miles; Bechhofer, 2009). It is based on the identification of similarities between systems of knowledge organization with the purpose of making them explicit to allow the sharing of data and technology through various applications. It uses RDF language and identifies concepts with URI that are documented with several types of notes. The concepts are linked together by hierarchies, associations and aggregates into conceptual schemes that can be mapped to the terms of other schemes.

Therefore, the model relates the nodes that represent the subject, predicate or property, and object, and form triples, that is, connects concepts following the premise that meaning is expressed in RDF (Berners-Lee

et al., 2001). The standard RDF concerning the formalization of vocabularies is still rarely mentioned in Brazilian Portuguese literature, except in the case of Boccato (2011), which clearly indicates the need for investment. The crucial point to allow triples to be consistent and to allow the codification of vocabularies that makes sense is the existence of a link between the concept and the term. The morphological coincidence is not sufficient to ensure interoperability.

Moreiro-González *et al*. (2012) point out the indications of the American National Information Standards Organization (ANSI/NISO) Z39.19:2005 and the British Standards Institution (BS) 8723-4:2007 for successful interoperability, considering that it depends on the similarity of concepts from different content areas/areas of knowledge, the use of different KOS's to index the contents from similar areas, the degree of specificity or granularity of documentary languages used, the search methods, the literary and organizational guarantee used in vocabulary development, and the purpose of different users for searching the databases (American National Information Standards Organization, 2005; British Standards Institution 2007).

Considering SKOS as an ontology, Pastor Sánchez *et al*. (2012) argue that, in comparison with other solutions, it is a simple and fast alternative. They criticize the way vocabulary data is published, propose the classification into categories to distinguish them, and observe that there is an excess of '*skosification*' of RDF datasets, since many of them cannot truly be characterized as SKOS.

To achieve interoperability standards, the formats are essential for the Semantic Web, as well as for the survival of earlier built vocabularies, ensuring their dissemination, sharing and expansion. Their achievement, however, cannot be restricted to a mechanical 'translation' under penalty of reproducing original problems of conceptual consistency on a larger scale. For this reason, it is essential to consider the recommendations of the ISO standards related to the subject.

Standards and recommendations

The pursuit of interoperability is what motivates the review of documentary standards for the development of controlled vocabularies and thesauri. The first thesauri standards (ISO 2788:1986 and ISO 5964:1985 and equivalents) did not consider Web environments (International Standard Organization, 1985, 1986). Yet ANSI/NISO Z39.19-2005 and BS 8723-4/5:2007/2008 introduced elements which sought interoperability (American National Information Standards Organization, 2005; British Standards Institution, 2007, 2008). The Guidelines for multilingual thesauri of the International Federation of Library Associations and Institutions (IFLA) (2009) and the proposals of SKOS by the W3C, which were made in 2004 and 2009 (World Wide Web Consortium, 2012), also influenced the revision of the standard, culminating in the recent publication of ISO 25964-1:2011/2012 - Thesauri and interoperability with other vocabularies, which is organized into two parts: pt. 1, Thesaurus for information retrieval; and pt. 2, Interoperability with other vocabularies.

An excellent review of the previous standards and recommendations and of the first part of the new standard was performed by Sánchez-Cuadrado *et al*. (2012). They highlight the changes related to the format of the thesauri (from paper to electronic support), the importance given to the concepts previously granted to the words, the abandonment of the differences between mono and multilingualism, the increased functionality through the adaptation to new user profiles, and the development of mapping mechanisms to allow interoperability. Considering this last aspect, they point out that in addition to the difficulty in finding the exact equivalences among thesauri, there are problems concerning their differences of quality, purpose and granularity level.

When analyzing the first part of the ISO 25964-1:2011 standard and without exhausting the list, other aspects and modifications were pointed out: the privilege of thesauri mentioning other types of vocabularies; the applicability to other resources beyond the text (sound, image movement, multimedia objects); the greater number of concepts related to the body of the standard; the treatment of different associative relationships, even when one thesaurus is used by speakers of different languages; the treatment of equivalence relations within the same language and in different languages; the amendment of the chapter on compound terms, which are now considered as complex terms; the chapter on faceted analysis; the improvement of symbology to indicate relations, allowing for differentiation among

types, the additional aspects (definitions, notes on the history of the term, categories); as well as the recommendations for thesauri management software etc. (International Standard Organization, 2011).

The second part of the standard (ISO 25964-2:2012) is dedicated to the mapping of the thesauri, including other vocabularies. Parallel to the symbolism of equivalence within the same vocabulary, it is recommended to achieve equivalence among languages comprising structural models for the mapping between vocabularies, the mapping types (equivalent, hierarchical, associative), the use of mappings for information retrieval, the treatment of the pre-coordination in the mappings, management, the display of mappings etc. Some chapters are also dedicated to the mapping of other forms of vocabularies, rather than thesauri. For each type of vocabulary, its characteristics, scope, role in information retrieval, relationships and semantic components in comparison with the thesauri described, followed by recommendations for mapping and examples that are not thesauri. The standard highlights the problems and limits of using interchange formats (MARC, Zthes, DD8723--5 and SKOS), indicating the need for adjustments. As for mapping, it is assumed that the vocabularies should be kept as separate entities that can be interconnected through their respective concepts, but the attachment of the standard registers, as an alternative, the idea of building a repository of terminological data from which different concepts, and vocabularies can be extracted. This is one possible approach that encourages cooperative work, but it must be noted that "To achieve this, we need a standardized data model" (International Standard Organization, 2012, p.98).

Transformation of the KOS into SKOS: conceptual problems

The codification of thesauri and other vocabularies in the formats for a Semantic Web like SKOS requires caution, given that the operation is not only instrumental and this is the moment when problems of original construction related to the explanation of concepts and their relations become evident. The origin of these problems is the failure to observe the principles of logical-semantic organization of the vocabularies. A major difficulty is the generalization made in many thesauri when using the Broader Term (BT) and the Narrower Term (NT), not only for logical relations that involve superordination and subordination (genus/species), but also for relations involving members (or whole/part). When we can speak of subordination per se, the hierarchy - in its logical sense -, is genus/species (property inheritance). Although the relations of genus/species and whole/part are different in most thesauri, the same symbols are used to indicate two types of hierarchy. The same occurs with the associative relations which, although they may be distinguished by the type of bond that is established between one concept and another, end up being represented by a single Related Term (TR) symbol. These relations should indicate the type of association at stake (process and consequence, activity and agent, action and patient etc.).

A more serious problem is the poor structuring of vocabularies resulting from the random organization of the terms. In the absence of explicit starting points (definitions), the conjunctions and disjunctions are inconsistent and concepts that do not relate logically or semantically are often merged within the hierarchies. Similarly, the formal equivalence between terms that are only morphologically coincident is problematic. This case illustrates the importance of the term-concept bond.

These examples are not sufficient to answer the questions that one faces in formalization, since the mere codification of KOS in SKOS does not solve interoperability problems. The conceptual issues of a documentary language must be resolved "elsewhere" (Sánchez-Jiménez; Gil-Urdiciain, 2007, p.552), namely in the sphere of the principles of its construction. Knowledge organization systems operate with concepts represented by words, not by mere formal labels. Ideological and cultural differences also mark the myriad forms of organization of conceptual systems and these issues cannot always be resolved given the original incompatibility of concepts. As noted by García Gutiérrez (1998), certain hierarchies operate with judgments, assessments and interpretations, and not necessarily with semantic criteria.

For García Gutierrez (1998), alternatives for reducing the ideological codification may be used by exploring the suggestions of associative relations once the re-dispatches give rise to the possibility of choices,

since it abandons the rigidity of the hierarchical organization of concepts. Or, as stated by Olson (2002, p.389):

> We can sail off the teleological evolutionary road (wherever that takes us) wandering the side roads to the accidental discovery. We can avoid the limitations of hierarchy illuminating different connections than those of minor categories within major categories.

For Olson (2002), overlapping systems allow the identification of 'border objects', which may be more appropriate to meet the needs of specific communities of practice. Several types of vocabularies can be linked to show different signifiers which have common meanings, without giving priority to a main vocabulary or one vocabulary to the detriment of another, establishing quasi--hierarchical relations.

Final Considerations

The formalization of vocabularies is essential for the functioning of the Semantic Web and interoperability. Yet the effort to codify existing vocabularies is an operation that cannot ignore the application of the principles of a logical-semantic organization, avoiding reduction to an instrumental operation. Knowledge organization systems are not neutral and do not respond to all purposes universally: the delimitation of a domain structure and the structuring of their concepts define their limits of application. As there are various forms of organization there is always a target audience for the information, and no vocabulary, as well-organized as it may be, is appropriate for all existing contexts and situations.

The possibilities opened by the initiatives to promote the reuse of vocabularies do not erase/invalid the aforementioned condition. But the mapping of vocabularies and its comparison increase the range of choices. Without denying the ramification as a means of structuring, it is possible to relativize the initial choices and choose the most suitable concepts to contemplate certain contexts. The Roget's Thesaurus principle that gave rise to the thesaurus documentary, which focuses on the association of ideas, can thus be recovered. Although it is organized in categories, its main focus is the word which triggers associations.

References

AMERICAN NATIONAL INFORMATION STANDARDS ORGANIZATION. *Z39.19*: guidelines for the construction, format and management of monolingual controlled vocabularies. Bethesda: NISO, 2005. Available from: <http://www.niso.org/kst/reports/standards?step=2&gid=&project_key=7cc9b583cb5a62e8c15d3099e0bb46bbae9cf38a>. Cited: July 12, 2012.

BERNERS-LEE, T. Linked data. *Design Issues*, 2006. Available from: <http://www.w3.org/DesignIssues/LinkedData.html>. Cited: June 12, 2012.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. *Scientific American*, 2001. Available from: <www3.unisa.it/uploads/3845/semantic_web-berners_lee.pdf> . Cited: Aug. 11, 2012.

BOCCATO, V.R.C. Os sistemas de organização do conhecimento nas perspectivas atuais das normas internacionais de construção. *InCID*: *Revista de Ciência da Informação e Documentação*, v.2, n.1, p.165-192, 2011.

BRITISH STANDARDS INSTITUTION. *BS 8723-4*: structured vocabularies for information retrieval guide - interoperability between vocabularies. London: British Standards Institution, 2007.

BRITISH STANDARDS INSTITUTION. *BS 8723-5*: structured vocabularies for information retrieval - guide: interoperation between vocabularies and other components of information storage and retrieval systems. London: British Standards Institution, 2008.

GARCÍA GUTIÉRREZ, A. *Principios de lenguaje epistemográfico*: la representación del conocimiento sobre Patrimonio Histórico Andaluz. Granada: Instituto Andaluz del Patrimonio Histórico, 1998.

HJORLAND, B. Knowledge organization systems (KOS). In: HJORLAND, B. *Lifeboat of knowledge organization*. 2008. Available from: <http://www.iva.dk/bh/lifeboat_ko/CONCEPTS/knowledge_organization_systems.htm>. Cited: July 6, 2012.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS. *Guidelines for multilingual thesauri*. The Hague: IFLA Headquarters, 2009. Available from: <http://archive.ifla.org/VII/s29/pubs/Profrep115.pdf>. Cited: June 15, 2012.

INTERNATIONAL STANDARD ORGANIZATION. *ISO 5964*: guide to establishment and development of multilingual thesauri. Geneve: International Standard Organization, 1985.

INTERNATIONAL STANDARD ORGANIZATION. *ISO 2788*: guidelines for the establishment and development of monolingual thesauri. Geneve: International Standard Organization, 1986.

INTERNATIONAL STANDARD ORGANIZATION. *ISO 25964*: thesauri and interoperability with other vocabularies. Part 1: thesauri for information retrieval. Geneve: International Standard Organization, 2011.

INTERNATIONAL STANDARD ORGANIZATION. *ISO 25964*: thesauri and interoperability with other vocabularies. Part 2: interoperability with other vocabularies. Geneve: International Standard Organization, 2012.

MÉNDEZ, E. *RDF: un modelo de metadatos flexible para las bibliotecas digitales del próximo milenio*. In: JORNADES CATALANES DE DOCUMENTACIÓ, 7., 1999, Barcelona. *Actas*... Barcelona: E-LIS, 1999. Available from: <http://eprints.rclis.org/12694/>. Cited: June 8, 2012.

MÉNDEZ, E.; GREENBERG, J. Linked data for open vocabularies and HIVE's global framework. *El Profesional de la Información*, v.21, n.3, p.236-244, 2012.

MILES, A.; BECHHOFER, S. (Ed.). *Simple knowledge organization system reference*: W3C recommendation. 2009. Available from: <http://www.w3.org/TR/2009/REC-skos-reference-20090818/>. Cited: July 7, 2012.

MOREIRO-GONZÁLEZ, J.A.; CUADRADO, S.S.; MORATO LARA, J. Mejora de la interoperabilidad semántica para la reutilización de contenidos mediante sistemas de organización del conocimiento. *Encontros Bibli*, v.17, n.33, p.46-58, 2012.

OLSON, H. Review article: classification and universality application and construct. *Semiotica*, v.139, n.1/4, p.377-391, 2002.

PASTOR SANCHEZ, J.A.; MARTÍNEZ MÉNDEZ, F.J.; RODRÍGUEZ MUÑOZ, J.V. Aplicación de SKOS para la interoperabilidad de vocabularios contorlados en el entorno de linked open data. *El Profesional de la Información*, v.21, n.3, p.245-253, 2012.

PESET, F.; FERRER-SAPENA, A.; SUBIRATS-COLL, I. Open data y linked data: su impacto en el área de bibliotecas y documentación. *El Profesional de la Información*, v.20, n.2, p.165-173, 2011.

SANCHEZ-CUADRADO, S.; COLMENERO-RUIZ, M.J.; MOREIRO GONZÁLEZ, J.A. Tesauros: estándares y recomendaciones. *El Profesional de la Información*, v.21, n.3, p.229-235, 2012.

SÁNCHEZ-JIMÉNEZ, R.; GIL-URDICIAIN, B. Lenguajes documentales y ontologías. *El Profesional de la Información*, v.16, n.6, p.551-560, 2007.

SOERGEL, D. *The representation of knowledge organization structure (KOS) data*: a multiplicity of standards. Roanoke: JCDL, 2001. Available from: <http://www.dsoergel.com/cvwelcome.htm#JournalArticles>. Cited: July 6, 2012.

WORLD WIDE WEB CONSORTIUM. *Simple knowledge organization system*. San Francisco: W3C, 2012. Available from: <http://www.w3.org/2004/02/skos/>. Cited: June 10, 2012.