Thesaurus and subject heading lists as Linked Data

Tesauros e listas de cabeçalhos de assuntos como Linked Data

Everton Rodrigues BARBOSA[1] 0000-0002-1111-5861

Moisés Lima DUTRA[1] 0000-0003-1000-5553

Angel Freddy GODOY VIERA[1] 0000-0001-6657-4734

Douglas Dyllon Jeronimo de MACEDO[1] 0000-0002-3237-4168

Abstract

Most libraries put a lot of effort into developing subject headings or thesauri, which are used to index and retrieve information. Nevertheless, in the library field, controlled vocabularies are associated to authority records as authority files. In order to become findable by search engines, these authority files should be modelled on semantic vocabularies. This research proposes an authority-record conversion process for publishing thesauri and subject headings as linked data, by using the Simple Knowledge Organization Systems data model. To this purpose, we undertook a bibliographic and documentary research on the World Wide Web Consortium recommendation guidelines, which were used to produce a set of procedures and technologies to support the conversion proposal. This research provides evidences that controlled vocabularies are an important resource for improving information retrieval on the web. The proposed conversion process works as a quick guide for controlled vocabulary integration and reuse among users and systems on the linked data environment. Although the proposal was originally intended for a library setting, it can be applied and tested in another type of institution, such as documentation centres, museums, or cultural heritage archives. It can also be used in other linked open data projects.

Keywords: Authority records. Controlled vocabularies. Semantic Web. Simple Knowledge Organization System.

Resumo

Grande parte das bibliotecas concentram esforços em desenvolver cabeçalhos de assuntos ou tesauros, os quais são usados para indexar e para recuperar informações. No entanto, no campo das bibliotecas, os vocabulários controlados são associados aos registros bibliográficos como arquivos de autoridade. Para se tornarem localizáveis pelos mecanismos de pesquisa, esses registros de autoridade devem ser modelados em vocabulários semânticos. Esta pesquisa propõe um processo de conversão de registros de autoridades para a publicação de tesauros e de cabeçalhos de assuntos como dados abertos conectados, utilizando o modelo de dados Simple Knowledge Organization Systems. Para tanto, realizou-se uma pesquisa bibliográfica e documental sobre as diretrizes e a recomendação do World Wide Web Consortium, as quais foram usadas para produzir um

[1] Universidade Federal de Santa Catarina, Centro de Ciências da Educação, Programa de Pós-Graduação em Ciência da Informação. R. Eng. Agronômico Andrei Cristian Ferreira, s/n., Trindade, 88040-900, Florianópolis, SC, Brasil. *Correspondência para*/Correspondence to: M. L. DUTRA. E-mail: <moises.dutra@ufsc.br>

Received on January 27, 2021, final version resubmitted on May 11, 2021 and approved on July 7, 2021.

Como citar este artigo/*How to cite this article*

Barbosa, E. R. *et al*. Thesaurus and subject heading lists as Linked Data. *Transinformação*, v. 33, e200077, 2021. https://doi.org/10.1590/2318-0889202133e200077

conjunto de procedimentos e de tecnologias para apoiar a proposta de conversão. Este trabalho fornece evidências de que os vocabulários controlados são um recurso importante para melhorar a recuperação de informações na web. O processo de conversão proposto funciona como um guia rápido para a integração e a reutilização dos vocabulários controlados entre usuários e sistemas no ambiente de dados abertos conectados. Embora a proposta tenha sido originalmente destinada à realidade das bibliotecas, pode ser aplicada e testada em instituições de natureza diversificada, como centros de documentação, museus ou arquivos. Ela também pode ser usada em outros projetos de dados abertos conectados.

Palavras-chave: *Registros de autoridade. Vocabulários controlados. Web Semântica. Simple Knowledge Organization System.*

Introduction

Knowledge Organization Systems (KOS) refer to a number of items (*e.g.*, thesaurus, subject headings, classification systems, taxonomies, ontologies, and folksonomies) used to represent and organise – in a systematic and structural way – a knowledge domain. KOS are considered an important tool for information and knowledge indexing, classification, and categorization. Which the main goal of it is to support information retrieval, especially in the context of digital technologies. Normally, information retrieval is based on keywords, however, it could be a problem to get wrong answers due to ambiguity of words. For this reason, semantic search is used to improve the accuracy of searches by understanding the intent of the user and the meaning of the terms in the searching sentence.

The web is a resource that facilitates access to information in different formats and data sources. Due to the amount of data and the occurrence of linked data indicatives, libraries must incorporate their authority records on semantic web environment. Nowadays, many projects aim to share data in a structured way to interact with the linked open data cloud, which currently gains space in government area, and even in cultural-heritage institutions (Scholz, 2017). A large number of cultural heritage projects are modelling their data on semantic web patterns. Examples include American Art Collaborative (2019), Europeana (2019), and Musical Instrument Museums Collections (2019), which provides access to data collections (*e.g*. books, music, artwork, etc.) of libraries, archives and museums of Europe and the United States of America. Despite the increased use of semantic web applications in the library field, two challenges remain: (i) How to map and convert authority data into semantic vocabularies? (ii) How to enrich these resources through semantic annotations? It is necessary to use standards to represent KOS on the web and support data interoperability among information systems (W3C, 2009a).

Simple Knowledge Organization System (SKOS) emerges as a data model for formalising the structure of KOS as machine-readable data schemas, in order to support representation, use, and interoperability of controlled vocabularies in linked data environments (Zoghlami; Kerherve; Gerbe, 2011). Some cultural-heritage institutions have published their experiences about the use of SKOS. It is worth mentioning that best practices for web data are being discussed by the W3C's Data on the Web Best Practices Working Group (W3C, 2017a). We understand, however, that there is still a lack of better practice studies on SKOS-based models, especially those aiming to assist information for professionals to publish their controlled vocabularies on the web.

There are a growing number of controlled vocabularies projects existent on the linked data context, especially thesauri, such as The UNESCO thesaurus (2018), EUROVOC (2018), AGROVOC (2018), UK Archival Thesaurus (2018), and Library of Congress Subject Headings (2018) (W3C, 2009a). Most libraries focus on the development of the subject headings or thesauri, and they are used to index and retrieve information. Nevertheless, in the library field, controlled vocabularies are associated to authority records. In order to become findable by search engines, these authority files should be modelled on semantic vocabularies. This research focuses on controlled vocabularies maintained by libraries, such as subject headings or thesauri, and aims to propose a set of guidelines to publish them by using linked data technologies. This research claims to answer the following question: How can one convert authority records controlled vocabularies and publish them on the web by using the SKOS data model? To answer this question, we undertook a bibliographical and documentary research on the best practices for managing web data, in order to propose a conversion and publishing process for SKOS-based controlled vocabularies.

Semantic Web and Linked Data

Semantic web technologies are used in cases where collaboration between organisations is critical, or when a large amount of data needs to be processed and integrated on the web (Molli; Breslin; Hutchison, 2016). Various examples of the application of these technologies exist and are used in a broad range of application areas, such as supply chain management, media management, data integration, web search, and e-commerce, to name a few. The semantic web provides several concepts, technologies, and functionalities, among which the following stand out: Resource Description Framework (RDF), eXtensible Markup Language (XML), SPARQL Protocol and RDF Query Language (SPARQL), and Web Ontology Language (OWL). These concepts and technologies are managed and defined as recommendations by the World Wide Web Consortium (W3C).

World Wide Web Consortium stands that web data must be available for process, retrieval, and reuse (Berners-Lee, 2009). This idea proposes that the semantic web is not only about providing data availability on the web, but also about establishing relationships among them. Linked data is related to the technical data interoperability and have been one of the key elements of the semantic web constitution (van Hooland; Verborgh, 2015).

From the initial proposition until today, linked data has represented an advance for semantic web projects and, consequently, optimised the information retrieval in most digital sources. Moreover, different types of KOS (including subject headings or thesauri) can be designed from SKOS-based data models and published as linked data. Next section presents some applications of KOS in semantic web contexts.

Thesaurus and Subject Heading Lists on the Semantic Web

Semantic web technologies require robust and flexible controlled vocabularies to facilitate the semantic search in digital resource collections. Traditionally, controlled vocabularies have been effective within controlled environments; *e.g.* the organisation of collections in libraries. However, due to the heterogeneity of data formats, resources and the structure of digital information, controlled vocabularies must be modelled into an ontological schema, which requires semantic web standards for modelling conceptual relationships (Ramalho, 2015).

In library and information science, thesauri and subject headings have been used to specify domains and improve information retrieval. The most known of them are the thesauri, considered a relevant tool to offer clear definitions of concepts and their mutual relations, and are used to express document subjects with restrictive controlled terms. Thesauri provide synonyms and simplified syntactic structures (Dodebei, 2002). According to Harpring (2015), thesauri stand out among KOS because it is traditionally used for indexing content or querying, and represents a robust and flexible tool to improve information retrieval systems.

In general, thesauri are a type of KOS where the concepts covering a knowledge domain are represented by terms in hierarchical, associative and equivalence levels (Zeng, 2008). Terms are structured according to the sort of relationship that could be represented by a Broader concept, their subdivisions in Narrower concepts, and their associations with other terms, in Related concepts.

Despite the importance of thesauri to knowledge organization, libraries commonly use subject headings to index their collections. Reitz (2004, p. 691) defines subject headings as "the most specific word or phrase that describes the subject, or one of the subjects, of a work, selected from a list of preferred terms (controlled vocabulary) and assigned as an added entry in the authority record to serve as an access point in the library catalog." The difference between thesauri and subject headings are related to the level of relationship of terms. While thesauri have three relationship levels, subject headings commonly are organized just into one hierarchical relationship of terms (Colepícolo *et al.*, 2006; Haider, 2020). Besides that, both controlled vocabularies have contributed to organise document collections in libraries.

The XML is a mark-up language that plays an important role in the exchange of a wide variety of data on the web and elsewhere. It defines a set of rules for encoding documents in both human and machine-readable

formats (W3C, 2016). XML was launched by the W3C in 1998 to represent web data, including bibliographic data. The benefits of the XML became clearer during the 1990s when the US Library of Congress developed Standard Generalized Markup Language for the conversion of bibliographic records in Machine Readable Cataloguing (MARC) format into the XML language (Library of Congress, 2008).

Large part of the thesauri are managed by thesaurus management systems and stored in relational databases, in which the standard output format for data is an XML or text file (W3C, 2005a). In library filed, subject headings or thesauri could be stored in authority databases as authority files, which have specific metadata standards for modelling authority records and may even be expressed in XML format. Some common standards are: MARC 21 Format for Authority Data and the Metadata Authority Description (MADS).

The MARC 21 is designed to be a carrier for information concerning the authorised forms of names, subjects, and subject subdivisions to be used in constructing access points in MARC records, the forms of these names, subjects, and subject subdivisions that should be used as references to the authorised forms, and the interrelationships among these forms. Despite the possibility of converting MARC 21 records into XML, the MARC 21 format possess some limitations to represent authority records. For this reason, the Metadata Object Description Schema (MODS) and the MADS were developed (Library of Congress, 1999).

The MADS is an XML schema for an authority element set that may be used to provide metadata about agents (people, organizations), events, and terms (topics, geographies, genres, etc.). MADS serves as a companion to MODS to provide metadata about the authoritative entities used in MODS descriptions (Library of Congress, 2017).

Some thesaurus may be structured in XML language or represented in OWL language and RDF data model. These formats are acceptable for publishing thesaurus or subject heading lists as linked data. However, SKOS data model is recommended by W3C best practices for publishing KOS on the web.

Beyond that, some major libraries already use the SKOS standard for authority data, such as the national libraries of France, Spain, Germany, the United Kingdom, as well as the Library of Congress'Virtual International Authority File, which combines multiple library authority files into a single name authority service.

Indeed, XML is an important language used to publish and share authority records in semantic web contexts. However, there are still difficulties to formally express controlled vocabularies on the web. Hence, it would be quite useful to expand thesauri or subject heading relationships by making them more adaptable and reusable in different domains. For that, it would be necessary to formalise the structure according to a set of axioms similar to ontologies, given that ontologies have greater expressive capacity for modelling information systems (Garcia-Torres; Pareja-Lora; Pradana-López, 2008). Similarly, W3C (2009a) stands that thesauri can take advantage of being used together with the OWL, since the conversion to a KOS representation based on formal logic may contribute to express and exchange knowledge about a specific domain. Matthews, Miles and Wilson (2001) believe that thesauri projects for the semantic web should involve other propositions like ontologies or knowledge maps (or even both combined) to establish a paradigm of thesauri creation under the conceptual logic. In this sense, W3C presents the SKOS to provide effective support for web-based data organisation, allowing traditional KOS to be expressed and shared as computer-readable data, making them interoperable with other semantic web applications (W3C, 2009a).

Simple Knowledge Organization System (SKOS) provides a standard way to represent KOS using RDF model. The RDF is a language that represents information about web resources, and possesses several serialisation formats, including the XML (W3C, 2009a). RDF contributes to the interoperability between information and description systems because it enables to express relationships between objects along with their properties and values. RDF is a standard model for metadata representation on the web, and it has features that facilitate data merging to express information about resources identified by International Resource Identifiers. The RDF structure is a collection of triples that represent a statement of a relationship between things, which is denoted by the nodes it connects. Each triple has three parts: subjects, predicates, and objects. While the subject and the object represent two related resources, the predicate (property) represents the nature of the relationship, formulated in a directional way (W3C, 2014a). The next section presents the methodological procedures used to construct the proposal.

Methodological Procedures

This is an applied research, which focuses on the solution for a practical problem. This research follows a bibliographic and documentary research to select academic papers and documents. This material was used as theoretical reference for supporting the guideline proposal.

A survey was carried out in the following databases: SCOPUS, Web of Science, and Library and Information Science Abstracts. The keywords "simple knowledge organization" AND "thesaurus" OR "subject headings" AND "workflow" OR "conversion process" were used. This refinement resulted in 83 related papers, from both journals and conference proceedings published in library and information science field. After a careful analysis, it was verified that only 14 papers were directly aligned with our proposal.

Some documents available on the W3C official website were also used: the SKOS Core Guide (W3C, 2005b); the SKOS Core Reference (W3C, 2009a); and the Quick Guide to Publishing Thesaurus on the semantic web (W3C, 2005c). These reference guides present recommendations for publishing KOS in the SKOS model. In addition, this proposal uses contributions taken from the following papers, listed in Table 1.

Next, the proposal of an authority-record conversion process for converting and publishing SKOS-based thesauri and subject-heading lists on the web is presented.

Results and Discussion

Our proposal comprises six main steps, represented in Figure 1: Data selection and extraction; Data mapping and conversion; RDF declaration; RDF/Turtle Serialisation; Open data Licensing; and Publication query. The six proposed steps are described below.

Data selection and extraction

The first step is to select controlled vocabularies from authority records, which are usually stored in authority databases. These records may be structured in MARC, MADS or even XML, RDF and OWL languages (Dunsire; Willer, 2011). Then it is necessary to analyse the format of the thesauri or subject headings in order to find out how their concepts are structured and how they are codified into the authority database (van Assem; Malaisé; Schreiber, 2006). Some authority database records must be extracted in XML format to be used during the step of data mapping and conversion. W3C (2005c, online) states that "[...] the standard output of thesaurus management systems is an XML or structured text format [...]", allowing records to be easily represented in RDF.

The use of XML, DTD, XML Schema and XSLT technologies to exchange authority data has been the subject of many papers, in which they are used to define structure and content of XML file our proposal focuses on authority databases and the library field. For those systems that work with other metadata formats, there is the possibility of extracting their records as non-XML files, such as PDF, DOC, or TXT. There are a number of tools to convert from these formats to XML (Leroi; Holland, 2010).

After that, a data mapping and conversion to SKOS data model is required. The main purpose of this step is to select a subject heading fragment, especially a specific domain group of concepts, in order to make an initial experiment. XML-based files are a prerequisite to construct RDF declaration links that provide semantic web perspective on authority records (W3C, 2014a).

Data mapping and conversion

The XML files extracted are used as an input to the mapping and conversion process of the authority records to the SKOS data model. The fact that both authority files and SKOS patterns use an XML-based syntax is a facilitator of

Table 1 – Studies used as a basis for the controlled vocabularies conversion process to SKOS.

Authors	Steps	Contributions
van Assem, Malaisé and Schreiber (2006) Dunsire and Willer (2011) Leroi and Holland (2010) Rudic and Surla (2009) W3C (2005c) W3C (2014a)	Selection and Extraction	Analysis and selection of records in bibliographic database. Extraction of authority records in XML format.
Harper (2006) Isaac and Tudhope (2015) Leroi and Holland (2010) Pastor-Sanchez, Martínez-Mendez and Rodriguez-Muñoz (2009) Pastor-Sánchez (2015) Summers, Isaac, Redding and Krech (2008) W3C (2009a)	Mapping	Metadata mapping of authority record formats to SKOS vocabulary properties. Recommendation for the use of Dublin Core metaproperties.
Bandholtz *et al.* (2010) Harper (2006) W3C (2009a)	Conversion	Conversion of XML records into SKOS/RDF.
Anibaldi *et al.* (2015) Isaac and Tudhope (2015) Leroi and Holland (2010) Pastor-Sánchez (2015) W3C (2005c) W3C (2009a) W3C (2014b) W3C (2017b)	Modelling	RDF declarations.
Anibaldi *et al.* (2015) Heather and Bizer (2011) W3C (2005c) W3C (2009a)	Serialisation	Serialisation syntax.
Korn and Oppenheim (2011) Creative Commons (2018)	Open Data Licences	Creative Commons licences to provide reuse of thesaurus datasets published as linked open data.
Anibaldi *et al.* (2015) Basharat, Abro, Arpinar, Rasheed (2016) W3C (2005c) W3C (2008) W3C (2013) W3C (2019)	Publication and Query	Linked open data publication and SPARQL endpoints.

Source: Elaborated by the authors (2020).

Note: RDF: Resource Description Framework; SKOS: Simple Knowledge Organization System; SPARQL: Protocol and RDF Query Language; XML: eXtensible Markup Language; W3C: World Wide Web Consortium.

such a process (Harper, 2006). Figure 2 illustrates the conversion that can be automatically done by using the eXtensible Stylesheet Language for Transformation (XSLT), as recommended by the W3C (2009a, online) and Harper (2006).

In order to use XLST transformations for converting XML files into their SKOS counterparts, the user has to know the structure of the terminology and be able to map the elements (Leroi; Holland, 2010; W3C, 2017b). The mapping of authority formats into SKOS provides a path to the interoperability between resources. Mapping fields, class or sub-class of MARC and MADS relates to SKOS classes and parent properties. For example, marc:datafieldtag="150" is a sub-property of skos:prefLabel, mads:Authority is a sub-class of skos:Concept.

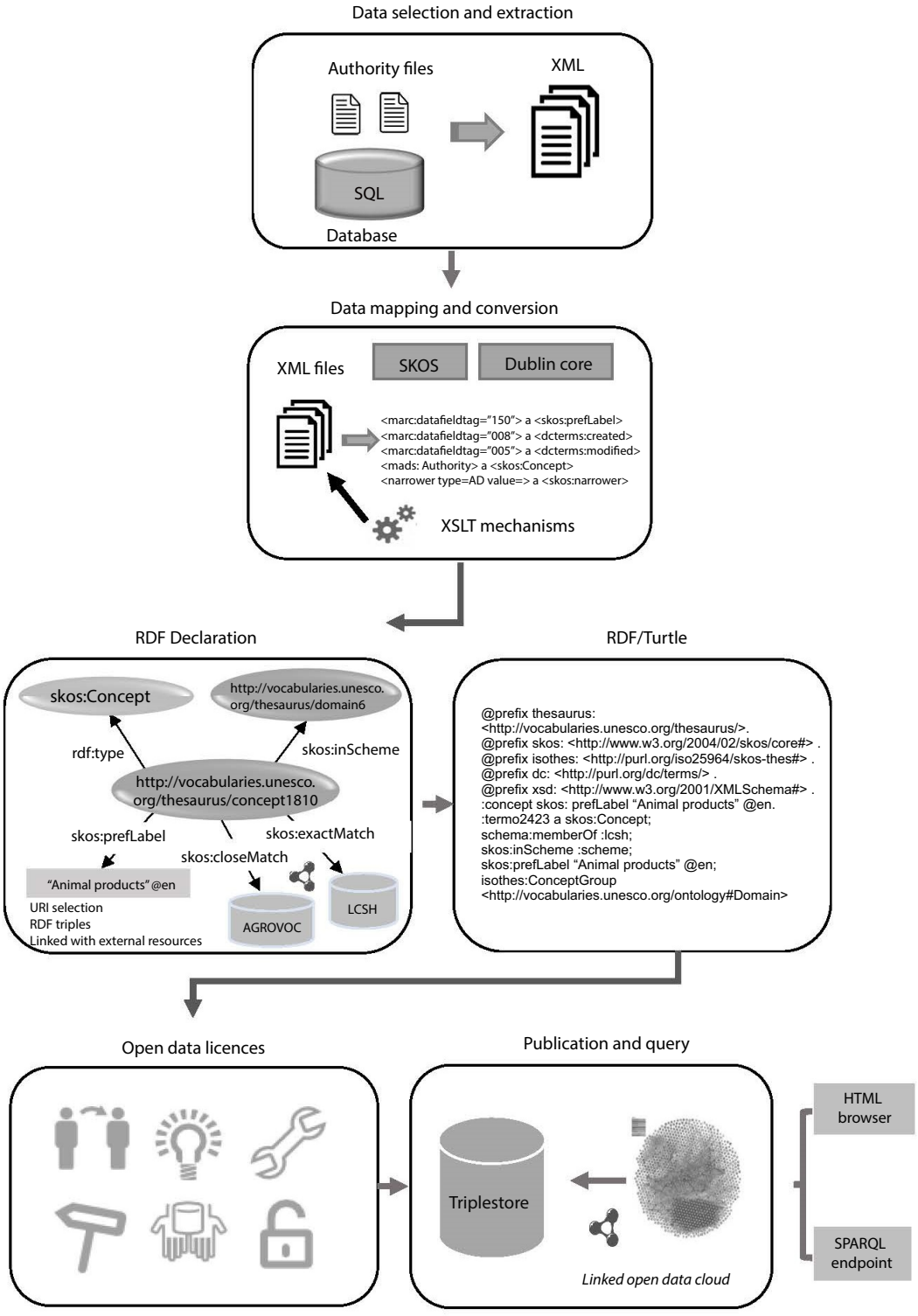

Figure 1 – General view of authority record conversion process for publishing thesauri and subject headings lists as linked data.

Source: Elaborated by the authors (2021).

Note: RDF: Resource Description Framework; SKOS: Simple Knowledge Organization System; SPARQL: Protocol and RDF Query Language; SQL: Structured Query Language; URI: Uniform Resource Identifier; XML: eXtensible Markup Language; XSLT: eXtensible Stylesheet Language for Transformation.

Authority records contain other features that need to be represented, such as: created or modified periods, page titles, and attribution licence categories. RDF can also be used to express metaproperties that do not possess equivalent tags in the SKOS vocabulary, such as website title, description, modification date, and copyright. Furthermore, a range of metadata standards can be used as an alternative to describe digital resources. They also provide interoperability between semantic web applications, including properties for publishing conceptual schemas on the web (Harper, 2006; W3C, 2005c). Although the SKOS vocabulary does not have properties to express this information, the RDF format allows other vocabularies, such as Dublin Core – mainly through properties illustrated in Figure 2: marc:datafieldtag="008" is a property of dcterms:created, and marc:datafieldtag="005" is a property of dcterms:modified – to be incorporated and combined with SKOS properties (Summers *et al.*, 2008).

The Dublin Core is used to represent complex documentation structures as a related resource description, such as creator, dates, modification dates, and some notes (Laporte; Mougenot; Garnier, 2012; Díaz-Corona *et al.,* 2019). SKOS core guide sets that "Both the Dublin Core element set and the Dublin Core terms vocabulary have a number of properties suitable for describing the basic properties of a concept scheme (*e.g.* dc:publisher, dcterms:audience, dcterms:issued etc.)" (W3C, 2004, online).

RDF Declaration

The RDF model is a powerful tool to semantically representation, publishing and interlinking web resources. Within the scope of KOS, RDF has been used to represent – among others – the concepts of thesauri or subject headings, modelled by using the SKOS vocabulary. To this purpose, each concept must be identified by its International Resource Identifier (IRI) (W3C, 2009b). Figure 3 presents an example of the "Animal Products" concept, linked to other terms in the UNESCO thesaurus. The RDF graph below illustrates linked terms by means of applying classes and properties taken from the SKOS vocabulary.

In order to convert hierarchical classifications to SKOS/RDF and edit the RDF triples, the user can use a SKOS management tool, for example: SKOSEd, SKOS2OWL, iQvoc (Bandholtz *et al*., 2010; W3C 2009a). RDF triples (W3C, 2009a) are used to model the concepts extracted from the thesauri or subject headings. In this part, we need to select the concept of URIs, fine-tune them into RDF triples and link them to external controlled vocabulary on the web (*e.g.* a concept in LCSH may be aligned to a concept in AGROVOC or UNESCO Thesaurus by using SKOS close and exact match properties: *skos:closeMatch* and *skos:exactMatch*). The use of SKOS specification allows linking a resource to other web resources by means of related-concept URIs (Anibaldi *et al*., 2015). Once the concepts are identified by URI/ IRI and the dataset is structured in RDF, it is possible to link them or combine them with other RDF documents. It allows thesauri datasets to be published as linked data. Furthermore, terms could be linked to other thesauri existent on the web, once they are published as SKOS models too (Isaac; Tudhope, 2015; Pastor-Sánchez, 2015).

Regarding the recommendations for publishing data in RDF format, W3C presents a set of documents on its official website, which includes: an introductory guide to describe RDF vocabularies and give an overview of some deployed RDF applications; an RDF data model description; an RDF technical specification; and RDF serialisation formats (W3C, 2014a, 2014b, 2014c). W3C (2005a, online) emphasizes that "[...] the simplest way to publish RDF data is to create one or more RDF documents containing their data and publish them on the Web via a normal HTTP server". In addition to that, RDF can be presented in a serialisation format.

Turtle Serialisation

Resource Description Framework (RDF) files need to be serialised (*i.e.,* encoded as a series of characters) according to any RDF syntax. Serialisation allows the triple store to be easily viewed and reused by other web applications (Anibaldi *et al*., 2015). Beyond RDF/XML serialisation syntax, there are other alternative RDF syntaxes for serialisation: Notation 3 (N3), Turtle, N-Triples, and the JavaScript Object Notation for Linked Data (JSON-LD) (Heath; Bizer, 2011).

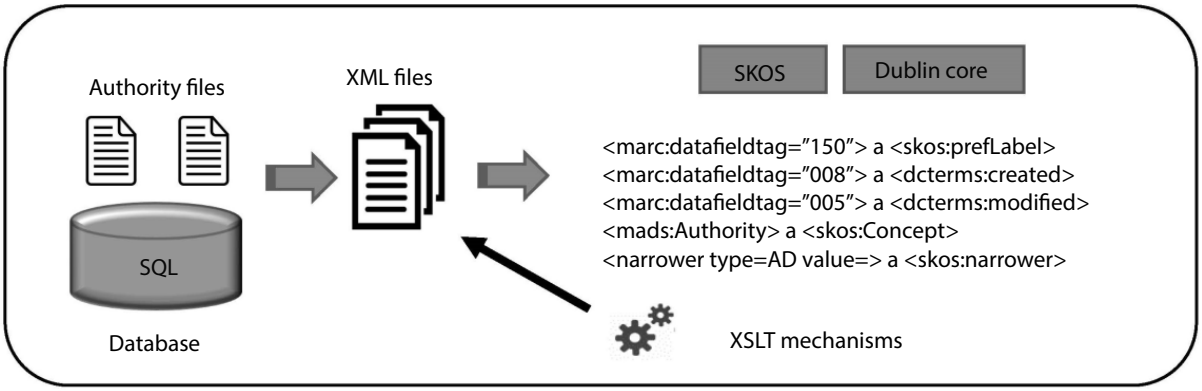

Figure 2 – Authority files selection, mapping and conversion to SKOS data model.

Source: Elaborated by the authors (2021).

Note: MADS: Metadata Authority Description; MARC: Machine Readable Cataloguing; SKOS: Simple Knowledge Organization System; SQL: Structured Query Language; XML: eXtensible Markup Language; XSLT: eXtensible Stylesheet Language for Transformation.

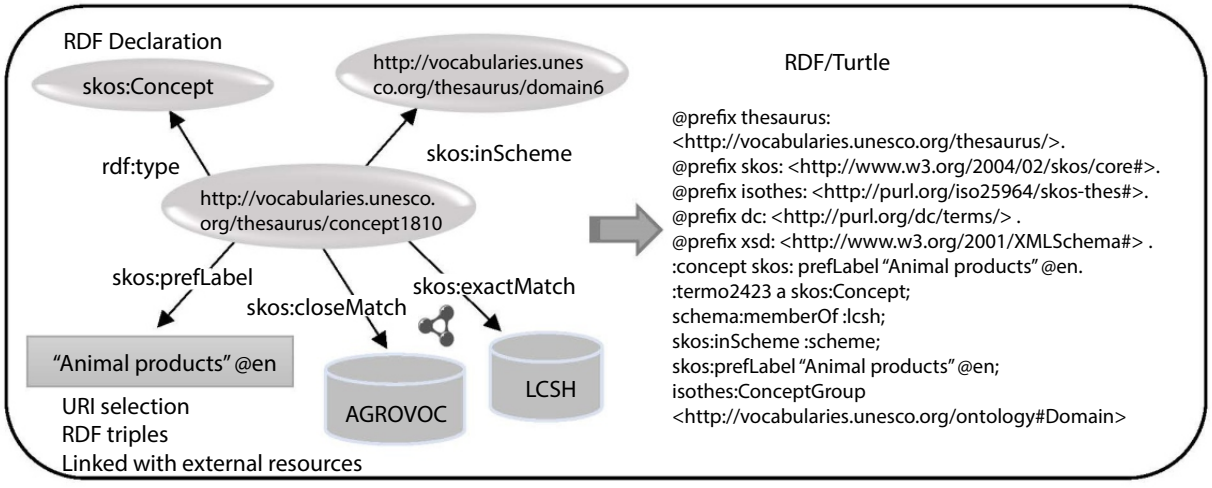

Figure 3 – RDF declaration and Turtle syntax.

Source: Elaborated by the authors (2021).

Note: AGROVOC: multilingual controlled vocabulary covering of the Food and Agriculture Organization of the United Nations; LCSH: Library of Congress Subject Headings; RDF: Resource Description Framework; SKOS: Simple Knowledge Organization System; URI: Uniform Resource Identifier.

The proposed guidelines suggest the use of the RDF/Turtle format, since it is a quite-simple computationally-processable and human-readable syntax (W3C, 2005c). RDF/Turtle provides a simple triple statement (Figure 3) in which the sequence of subject-predicate-object terms are separated by whitespaces. Each triple statement ends with a period, and it is also possible to use URIs as triple elements. Besides that, Turtle syntax provides a compatibility with the N-Triple format and the triple pattern syntax of the SPARQL W3C Recommendation (W3C, 2014b).

Open data licences

Creative Commons is one of the easiest attributions licences to use. It is fast becoming very popular in web applications that publish open content. One of the reasons behind the largely use of Creative Commons by linked open data applications is the fact that it is easily understood by humans. Figure 1 presents a set of seals of the Creative

Commons to setting permissions regarding the access and use of the published content. Among these, it can be highlighted the CC0 (Universal - Dedication to the Public Domain), which is ideal for publishing linked open data in the public domain, since it suggests the renunciation of all copyright (Korn; Oppenheim, 2011).

Publication and query

Once represented in SKOS data model and stored in RDF datasets, controlled vocabulary can be put available on the web for the use of other semantic web applications. Once they are publicly available, they can be merged with other RDF datasets, as well as connected to the linked open data cloud, as shown in Figure 1. According to Summers *et al.* (2008), a SPARQL Endpoint allows users to retrieve concepts more effectively, because it does not need to download and index the dataset altogether before exploring it. An SPARQL endpoint is an interface through which users or clients can query an RDF dataset through a web service. The query can be distributed to several other SPARQL Endpoint Query Services as federated query. The growing number of SPARQL query services offers data consumers an opportunity to merge data distributed over the Web. Thus, SPARQL Endpoint services have been widely used in linked data projects (W3C, 2013). A SPARQL service would bring practicality on the data web search (W3C, 2008).

A huge number of SPARQL technologies exist to deploy large triple stores. Moreover, a variety of systems currently use them to provide SPARQL queries on relational data, *e.g.* SPARQLer, D2RQ, and Open Link Virtuoso (W3C, 2019). These technologies allow RDF data to become available as linked data, providing SPARQL queries through a SPARQL Endpoint (Basharat *et al*., 2016; Anibaldi *et al*., 2015). Through an SPARQL Endpoint, the query can be distributed to several servers, facilitating data manipulation and data operation by using RDF language sentences, or triple standards. Theses sentences retrieve a structured set of data, both human and machine readable.

Conclusion and future works

The theoretical supports used for building our proposal allowed us to better understand the importance of KOS in information retrieval, especially through a semantic web perspective. Although traditionally built under rigid lexical frameworks, the development of W3C standards, technologies, and recommendations has enabled thesauri and subject heading to expand their value in linked open data environments. It is possible to notice that some institutions around the world (*e.g.* UNESCO, *Food and Agriculture Organization*, Europe Union, UK National Archives) are publishing their thesauri in linked open data format, by using the SKOS vocabulary. However, there is still a lack of information on how they are executing this task. This is not a widespread scenario, nevertheless. Perhaps the reason for the low participation of some institutions in this kind of work is due to the lack of well-defined institutional policies to publish open data, as well as the lack of reference models and best-practice guides. This initial realisation motivated this research. Subsequently, we sought to identify semantic web requirements and technologies to build our proposal.

According to the background material used to construct the proposal, we noticed that some challenges should be tackled during the two first steps of the proposal: one (data selection and extraction) and two (data mapping and conversion). Extracting a concept domain can be a problematic process if the authority records possess textual notes with several kinds of knowledge embedded. Nevertheless, some level of concept relationship has to be adapted with alternative SKOS ontologies and extensions, specially to enable alignment and interoperability with other web vocabularies. It should also be said that despite the possibility of extracting and converting authority records, the process of mapping them to SKOS should also use alternative vocabularies to specify non-bibliographic data present in these records. This is so because the SKOS vocabulary only covers the representation of concepts and their relationships. Creation and modification dates and license information are not covered by the SKOS vocabulary.

Yet, it is important to remind that this proposal is based only on W3C recommendations and some background material, such as journal and conference papers. The literature review gave us a theoretical perspective to achieve an accuracy path towards the publication of SKOS-based controlled vocabularies. In order to validate this proposal, the next step is to carry out a practical test. Despite the proposal is concerned to a library's scenario, it can also be

applied and tested in some different cultural heritage institutions, such as museums or archives. Furthermore, it can also be used in other linked open data projects, especially those working with widely disseminated vocabulary, shared by the scientific community. Finally, future works should invest in bringing together best practices for web data, and in providing mechanisms for automatically converting data into SKOS.

Contributors

E. R. BARBOSA was responsible for the research design, literature analysis, systematization and analysis of results and final writing. M. L. DUTRA, A. F. GODOY VIERA and D. D. J. MACEDO were responsible for the research design and writing review. All the authors are responsible for the approval of the final version.

References

Agrovoc. *AGROVOC Linked open data*. Rome: Food and Agriculture Organization, 2018. Available from: http://aims.fao.org/standards/agrovoc/linked-data. Cited: Oct. 29, 2020.

American Art Collaborative. *Linked open data initiative*. Washington: AAC, 2019. Available from: http://americanartcollaborative.org/. Cited: Oct. 29, 2020.

Anibaldi, S. *et al*. Migrating bibliographic datasets to the semantic web: the AGRIS case. *Semantic Web,* v. 6, p. 113-120, 2015. Available from: https://content.iospress.com/articles/semantic-web/sw128 .Cited: Aug. 27, 2020.

Bandholtz, T. *et al*. iQvoc – Open Source SKOS (XL) Maintenance and Publishing Tool. *In*: Workshop on Scripting and Development for the Semantic Web, 6., 2010, Heraklion. *Proceedings Online* […]. Heraklion; 2010. Available from: http://ceur-ws.org/Vol-699/Paper2.pdf. Cited: Aug. 27, 2020.

Basharat, A. *et al.* Semantic hadith: leveraging Linked Data opportunities for Islamic knowledge. *In*: Workshop on Linked Data on the Web (LDOW2016), 2016, Montreal. *Proceedings Online* […]. Montreal; 2016. Available from: http://events.linkeddata.org/ldow2016/papers/LDOW2016_paper_06.pdf. Cited: Aug. 27, 2020.

Berners-Lee, T. *Linked Data - design issues*. [*S. l.*]: W3C, 2009. Available from: http://www.w3.org/DesignIssues/LinkedData.html. Cited: Aug 27, 2020.

Colepícolo, E. *et al*. MeSH: de cabeçalho de assunto a tesauro. *In*: Congresso Brasileiro de Informática em Saúde, 10., 2006. *Anais eletrônicos* [...]. Florianópolis: SBIS, 2006. Available from: https://www.researchgate.net/publication/228885645. Cited: Aug. 27, 2020.

Creative Commons. *About the licences*. Mountain View: Creative Commons, [2018]. Available from: https://creativecommons.org/licenses/?lang=en. Cited: Oct. 29, 2020.

Díaz-Corona, D. *et al.* Profiling of knowledge organisation systems for the annotation of Linked Data cultural resources. *Information Systems*, v. 84, p. 17-28, 2019. Doi: https://doi.org/10.1016/j.is.2019.04.008.

Dodebei, V. L. D. *Tesauro*: linguagem de representação da memória documentária. Niterói: Intertexto, 2002.

Dunsire, G.; Willer, M. Standard library metadata models and structures for the semantic web. *Library Hi Tech News*, v. 28, n. 3, p. 1-12, 2011. Doi: https://doi.org/10.1108/07419051111145118.

Europeana. *Europeana Collections.* Hague: Connecting Europe Facility, 2019. Available from: https://www.europeana.eu/. Cited: Aug. 27, 2020.

Eurovoc. *European Union multilingual thesaurus.* Luxemburgo: Eurovoc, 2018. Available from: https://data.europa.eu/euodp/es/data/dataset/eurovoc. Cited: Oct 29, 2020.

García-Torres, A.; Pareja-Lora, A.; Pradana-López, D. Reutilización de tesauros: el documentalista frente al reto de la Web semántica. *El Profesional de la Información*, v. 17, n. 1, p. 8-21, 2008. Doi: https://doi.org/10.3145/epi.2008.ene.02.

Haider, S. Vocabulary control. *Librarianship Studies & Information Technology.* [Deli], Mar. 16, 2020. Available from: https://www.librarianshipstudies.com/2017/03/vocabulary-control.html. Cited: Aug 27, 2020.

Harper, C. A. Encoding Library of Congress subject headings in SKOS: authority control for semantic web. *In*: International Conference on Dublin Core and Metadata Applications, 2006. *Proceedings Online* […]. Colima: Dublin Core Metadata Initiative, 2006. Available from: http://dcpapers.dublincore.org/pubs/article/view/842. Cited: Aug. 27, 2020.

Harpring, P. Controlled vocabularies in Context. *In*: Harpring, P. *Introduction to controlled vocabularies*: featuring the Getty Vocabularies. Los Angelis: Getty Research Institute, 2015.

Heath, T.; Bizer, C. *Linked Data*: evolving the web into a global data space. Williston: Morgan & Claypool, 2011. Available from: http://linkeddatabook.com/editions/1.0/. Cited: Aug. 27, 2020.

Isaac, A.; Tudhope, D. S. [*ISO-THES*]. 2015. Available from: http://pub.tenforce.com/schemas/iso25964/skos-thes. Cited: Aug. 27, 2020.

Korn, K.; Oppenheim, C. *Licensing Open Data*: a practical guide. United Kingdom: Discovery Programme, 2011. Available from: http://discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf. Cited: Aug. 27, 2020.

Laporte, M. A.; Mougenot, I.; Garnier, E. ThesauForm-Traits: a web based collaborative tool to develop a thesaurus for plant

functional diversity research. *Ecological Informatics*, v. 11, p. 34-44, 2012. Doi: http://dx.doi.org/10.1016/j.ecoinf.2012.04.004.

Leroi, M.-V.; Holland, J. *Guidelines for mapping into SKOS, dealing with translations*. Roma: Athena, 2010. Available from: https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/ATHENA/Deliverables/D4.2_Guidelines%20for%20mapping%20into%20SKOS.pdf. Cited: Nov. 3, 2020.

Library of Congress**.** *MARC in XML*. Washington: Library of Congress, 2008. Available from: https://www.loc.gov/marc/marcxml.html. Cited: Nov. 3, 2020.

Library of Congress. *MARC 21 Format for Authority Data*. Washington: Library of Congress, 1999. Available from: https://www.loc.gov/marc/authority/. Cited: Nov. 3, 2020.

Library of Congress. *Metadata Authority Description Schema:* official Web Site. Washington: Library of Congress, 2017. Available from: http://www.loc.gov/standards/mads/. Cited: Nov. 3, 2020.

Library of Congress. *Library of Congress Subject Headings*. Washington: Library of Congress, 2018. Available from: http://id.loc.gov/authorities/subjects.html. Cited: Nov. 3, 2020.

Matthews, B.; Miles, A.; Wilson, M. *Modelling thesauri for the semantic Web*. Semantic Web Advanced Development for Europe (SWAD-Europe). [2001]. Available from: http://www.webcitation.org/5m2lmCyQY. Cited: Aug. 27, 2020.

Molli, P.; Bleslin, J.; Hutchison, D. *Semantic Web Collaborative Spaces*. Switzerland: Spring, 2016. Doi: http://doi.org/10.1007/978-3-319-32667-2.

Musical Instrument Museums Collections. *Museum Collections*. Paris: MIMO, 2019. Available from: http://www.mimo-db.eu/. Cited: Nov. 3, 2020.

Pastor-Sánchez, J. Proposal To Represent the Unesco Thesaurus for the Semantic Web Applying ISO-25964. *Brazilian Journal of Information Science, v.* 9, n. 2, p. 1-8, 2015. Available from: https://doi.org/10.36311/1981-1640.2015.v9n2.01.p1. Cited: Aug. 27, 2020.

Pastor-Sanchez, J. A.; Martínez-Mendez, F. J.; Rodríguez-Muñoz, J. V. Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives. *Information Research*, v. 14, n. 4, p. 1-15, 2009. Available from: http://www.informationr.net/ir/14-4/paper422.html. Cited: Aug. 27, 2020.

Ramalho, R. A. S. Análise do modelo de dados SKOS: Sistema de Organização do Conhecimento Simples para a Web. *Informação & Tecnologia,* v. 2, n. 1, p. 66-79, 2015. Available from: http://periodicos.ufpb.br/ojs/index.php/itec/article/view/25995. Cited: Aug. 27, 2020.

Rudic, G.; Surla, D. Conversion of bibliographic records to MARC 21 format. *The Electronic Library*, v. 27, n. 6, p. 950-67, 2009. Doi: http://dx.doi.org/10.1108/02640470911004057.

Reitz, J. M. *Online dictionary for library and information science*. London: Libraries Unlimited, 2004. Available from: https://www.abc-clio.com/ODLIS/odlis_s.aspx#subjectheading. Cited: Nov. 3, 2020.

Scholz, H. *Europeana publishing guide v1.5*: a guide to the metadata and content requirements for data partners publishing their collections in Europeana. Haia: Europeana Fundation, 2017. Available from: https://pro.europeana.eu/post/publication-policy. Cited: Nov. 3, 2020.

Summers, E. *et al.* LCSH, SKOS and Linked Data. *In*: International Conference on Dublin Core and Metadata Applications, 2008. *Proceedings Online* […] Berlin, 2008. Available from: http://dcpapers.dublincore.org/pubs/article/view/916. Cited: Aug. 27, 2020.

Unesco Thesaurus. *Thesaurus*. Paris: UNESCO, 2018. Available from: http://vocabularies.unesco.org/browser/thesaurus/en/. Cited: Aug. 27, 2020.

UK Archival Thesaurus. *Thesaurus.* London: The National Archives, 2018. Available from: https://ukat.aim25.com/thesaurus/. Cited: Aug. 27, 2020.

van Assem, M.; Malaisé, V. M.; Schreiber, G. A method to convert Thesauri to SKOS. *In*: Sure, Y.; Domingue, J. (ed.). *The Semantic Web*: research and applications. Berlin: Springer, 2006. (Lecture Notes in Computer Science, v. 4011). Available from: https://link.springer.com/chapter/10.1007/11762256_10. Cited: Aug. 27, 2020.

van Hooland, S.; Verborgh, R. *Linked Data for libraries, archives and museums*: how to clean, link and publish your metadata. London: Facet, 2015.

W3C. *SKOS Core Guide*. Cambridge: World Wide Web Consortium, 2005b. Available at: https://www.w3.org/TR/swbp-skos-core-guide/. Cited: Nov. 3, 2020.

W3C. *SKOS Core Guide*. Cambridge: World Wide Web Consortium, 2004. Available from: http://www.w3.org/2004/02/skos/core/guide/2004-10-22/. Cited: Nov. 3, 2020.

W3C**.** *SKOS Core Vocabulary*. Cambridge: World Wide Web Consortium, 2005a. Available from: https://www.w3.org/TR/2005/WD-swbp-skos-core-spec-20051102/. Cited: Nov. 3, 2020.

W3C. *Quick Guide to publishing a thesaurus on the semantic web*. Cambridge: World Wide Web Consortium, 2005c. Available from: https://www.w3.org/TR/swbp-thesaurus-pubguide/. Cited: Nov. 3, 2020.

W3C. *SPARQL Query LangUage for RDF*. Cambridge: World Wide Web Consortium, 2008. Available from: https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/. Cited: Nov. 3, 2020.

W3C. *SKOS Simples Knowledge Organization System Reference*: W3C Recommendation. Cambridge: World Wide Web Consortium, 2009a. Available from: http://www.w3.org/TR/skos-reference/. Cited: Nov. 3, 2020.

W3C. *SKOS Simples Knowledge Organization System Primer*: W3C Recommendation. Cambridge: World Wide Web Consortium, 2009b. Available from: https://www.w3.org/TR/skos-primer/. Cited: Nov. 3, 2020.

W3C. *SPARQL 1.1 Federated Query***.** Cambridge: World Wide Web Consortium, 2013. Available from: https://www.w3.org/TR/sparql11-federated-query/. Cited: Nov. 3, 2020.

W3C. *RDF 1.1 Primer*. Cambridge: World Wide Web Consortium, 2014a. Available from: https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/. Cited: Nov. 3, 2020.

W3C. *RDF 1.1 concepts and abstract syntax*. Cambridge: World Wide Web Consortium, 2014b. Available from: https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/. Cited: Nov. 3, 2020.

W3C. *RDF 1.1 Turtle*. Cambridge: World Wide Web Consortium, 2014c. Available from: https://www.w3.org/TR/turtle/#simple-triples. Cited: Nov. 3, 2020.

W3C. *Extensible Markup Language (XML)*. Cambridge: World Wide Web Consortium, 2016. Available from: http://www.w3.org/XML. Cited: Nov. 3, 2020.

W3C. *Data on the web best practices*. Cambridge: World Wide Web Consortium, 2017a. Available from: https://www.w3.org/TR/dwbp/. Cited: Nov. 3, 2020.

W3C. *XSL Transformations (XSLT) Version 3.0*. Cambridge: World Wide Web Consortium, 2017b. Available from: https://www.w3.org/TR/2017/REC-xslt-30-20170608/#initiating. Cited: Nov. 3, 2020.

W3C. *LargeTripleStores*. Cambridge: World Wide Web Consortium, 2019. Available from: https://www.w3.org/wiki/LargeTripleStores. Cited: Nov. 3, 2020.

Zeng, M. L. Knowledge Organization Systems (KOS). *Knowledge Organization*, v. 35, n. 2-3, p. 160-182, 2008. Available from: http://onlinelibrary.wiley.com/doi/10.1002/meet.145044019/full. Cited: Nov. 3, 2020.

Zoghlami, K.; Kerhervé, B.; Gerbé, O. Using a SKOS engine to create, share and transfer terminology data sets. *In*: The International Conference on Signal Image Technology & Internet-Based Systems, 7., 2011. *Proceedings Online* [...]. Los Alamitos: IEEE Computing Society, 2011. Available from: https://ieeexplore.ieee.org/document/6120628/. Cited: Nov. 3, 2020.