XX ENANCIB
21 a 25 Outubro/2019 – Florianópolis
A Ciência da Informação e a era da Ciência de Dados

ISSN 2177-3688
GT-2 – Organização e Representação do Conhecimento

ANÁLISE DE PARÂMETROS PARA AVALIAÇÃO AUTOMÁTICA DE VOCABULÁRIOS EM SKOS

ANALYSIS OF PARAMETER FOR EVALUATION AUTOMATIC OF VOCABULARIES IN SKOS

Janailton Lopes Sousa - Universidade Federal de São Carlos
Rogério Aparecido Sá Ramalho - Universidade Federal de São Carlos

Modalidade: Trabalho Completo

Resumo: Os vocabulários controlados destinam-se a padronização de termos representativos utilizados em domínios do conhecimento, além disso possibilitam estabelecer relações entre esses termos, o *Simple Knowledge Organization System* (SKOS) permite representar vocabulários controlados de modo simplificado, por meio dele é possível manipular, contextualizar dados e vincular conceitos, inferindo uma ordem semântica interpretada por máquinas, como é o caso do *Linked Data*. O objetivo deste trabalho é apresentar uma análise teórica sobre parâmetros de avaliação automática para vocabulários codificados em SKOS encontrados na literatura, como parte do resultado de pesquisa sobre o padrão SKOS, atualmente fomentada pela FAPESP e pelo CNPq. A metodologia utilizada fundamenta-se na pesquisa bibliográfica e revisão de literatura, possui uma abordagem exploratória e qualitativa. Como resultado é apresentada uma sistematização categórica dos parâmetros identificados na literatura que possam ser utilizados na criação de processos de avaliação para vocabulários em SKOS, contribuindo para o desenvolvimento de ferramentas e o fortalecimento de diretrizes que possam nortear a avaliação de Sistemas de Organização do Conhecimento (SOC) em ambientes digitais.

Palavras-Chave: SOC; Vocabulários controlados; SKOS; Linked Data.

Abstract: The Controlled vocabularies are intended to standardize representative terms used in knowledge domains, and in addition to establishing relationships between these terms, the Simple Knowledge Organization System (SKOS) allows to represent controlled vocabularies in a simplified way, through which it is possible to manipulate, contextualize data and link concepts, inferring a semantic order interpreted by machines, such as Linked Data. The aim of this paper is to present a theoretical analysis on automatic evaluation parameters for SKOS-coded vocabularies found in the literature, as part of the research results on the SKOS standard, currently promoted by FAPESP and CNPq. The methodology used was bibliographic research and literature review, has an exploratory and qualitative approach. As a result, a categorical systematization of the parameters identified in the literature that can be used inthe creation of assessment processes for SKOS vocabularies is presented, contributing to the development of tools and the strengthening of guidelines that can guide the assessment of Knowledge Organization Systems (KOS) in digital environments.

Keywords: KOS; Controlled vocabularies; SKOS; Linked Data.

1 INTRODUÇÃO

O uso de ferramentas tecnológicas tem impulsionado a publicação de vocabulários controlados, devido a inserção de diferentes padrões que permitem a migração destes Sistemas de Organização do Conhecimento (SOC) em para ambientes dinâmicos como a web. O *Simple Knowledge Organization System* (SKOS) é um desses padrões, que se caracteriza como uma aplicação do *Resource Description Framework* (RDF), pois permite que conceitos compostos publicados na web sejam vinculados e formem esquemas conceituais. Cada conceito é apontado por uma *Uniform Resource Identifier* (URI) e etiquetados por uma *tag* ou *string* em linguagem natural.

O SKOS fornece três propriedades para anexar etiquetas de recursos conceituais, que são: o <*skos:prefLabel*> que possibilita atribuir uma etiqueta preferida a um recurso, o <*skos:altLabel*> que permite atribuir uma etiqueta alternativa. E o <*skos:hiddenLabel*> um etiqueta léxica acessível a aplicativos que executam operações de indexação e pesquisa baseadas em texto, mas não se torna visível. As etiquetas ocultas podem, por exemplo, ser usados para incluir variantes com erros ortográficos de outras etiquetas lexicais. (ISAAC; SUMMERS, 2009)

Para estabelecer relacionamentos entre os termos são utilizadas três propriedades. Para relações hierárquicas, amplas e estreitas, utiliza-se respectivamente <*skos:broader*> e <*skos:narrower*> que permitem a representação de links hierárquicos, como a relação entre um gênero e suas espécies mais específicas, ou, dependendo de interpretações, a relação entre um todo e suas partes. (ISAAC; SUMMERS 2009). As relações associativas utilizam <*skos:related*>, para a representação de links associativos (não hierárquicos), como a relação entre um tipo de evento e uma categoria de entidades, que normalmente participam dele. (ISAAC; SUMMERS, 2009).

Considerando as características elementares do SKOS, a identificação de parâmetros que possam subsidiar a avaliação de vocabulários disponíveis em SKOS tornam-se mais viáveis. O objetivo deste trabalho é apresentar uma análise teórica sobre parâmetros de avaliação automática para vocabulários em SKOS, como parte do resultado de uma pesquisa sobre SKOS fomentada pela FAPESP e pelo CNPq. A metodologia utilizada foi a pesquisa bibliográfica e a revisão de literatura, possui uma abordagem exploratória e qualitativa, a discussão centra-se nos vocabulários controlados expressos em SKOS.

Os parâmetros apresentados são resultados da análise dos trabalhos realizados por

Suominen e Hyvönen (2012); Mader, Haslhofer e Isaac (2012); Suominen e Mader (2014), que por meio de análises automáticas de diferentes vocabulários controlados conseguiram identificar erros comuns de vocabulários em SKOS. A partir de uma análise teórica são apresentadas algumas características dos vocabulários controlados e a importância do processo de avaliação para esse tipo de SOC, especificamente àqueles que estão codificados em SKOS.

2 VOCABULÁRIOS CONTROLADOS

A linguagem natural é carregada por diversos elementos, que necessitam da cognição humana para torná-la possível de se entender, figuras de linguagem, sinonímia e homonímia são exemplos fundamentais para visualizar esta questão. Quando se trata de vocabulários controlados há uma tentativa de minimizar os vícios da linguagem natural e objetivá-la na medida do possível. “Um vocabulário controlado é essencialmente uma lista de termos autorizados para controlar sinônimos, diferenciar homógrafos e estabelecer as relações hierárquicas e não-hierárquicas” (LANCASTER, 2004, p. 19).

Os vocabulários controlados destinam-se a padronização de termos representativos utilizados em determinados domínios do conhecimento, além disso possibilitam estabelecer relações entre vários termos. Existem diversos tipos de vocabulários controlados, como cabeçalhos de assunto, glossários, taxonomias, tesauros, etc. Cada vocabulário possui características que podem se adaptar a diferentes necessidades.

Os Sistemas de Organização do Conhecimento expressos na forma de vocabulários controlados podem representar tanto estruturas simples como sistemas complexas. Um vocabulário controlado pode ser organizado de acordo com o grau de controle introduzido (da linguagem natural à linguagem controlada) e a força de sua estrutura semântica (de fracamente estruturada a fortemente estruturada), correspondendo às principais funções de SOC (NKOS, 2000).

Cada tipo de vocabulário cumpre uma função diferente, que pode limitar-se apenas a eliminação de ambiguidade no caso dos glossários, ou cumprir esta função, acrescida do controle de sinônimos, estabelecimento de relações hierárquicas e associativas no caso dos tesauros. De acordo com a ANSI/NISO Z39.19 (2010) o processo de organizar uma lista de termos objetiva:

a) indicar dois ou mais termos sinônimos que são autorizados para uso;
b) distinguir homógrafos;
c) indicar hierarquias e relações associativas entre os termos;

O processo de elaboração de um vocabulário controlado é pautado por várias etapas que podem demandar muito tempo, pois o simples fato de tentar controlar a linguagem exige um constante acompanhamento das mudanças implementadas, que pode ser desde a inserção de um novo termo, como a inclusão de novos suportes tecnológicos para impulsionar o eficiência de um vocabulário.

Tecnologias como os vocabulários vinculados têm se beneficiado da fundamentação teórica por trás dos SOC em seu processo de publicação, pois permitem a conexão semântica entre termos que podem representar conceitos por meio de uma série de relacionamentos, facilitando a convergência interoperável desses vocabulários. Sobre este aspecto os SOC oferecem os subsídios necessários para o entendimento das tecnologias semânticas que incorporam os vocabulários em SKOS.

2.1 Vocabulários em SKOS

Um SOC pode ser "fundamentado em termos" ou “conceitos” dependendo de como explicitamente pretende representar essas estruturas conceituais (BAKER et al, 2013). Notadamente os métodos manuais de organização são enfatizados devido sua tradicional usabilidade e adequação as demandas existentes. No entanto diversas iniciativas tecnológicas foram desenvolvidas ao longo do tempo, para otimizar estes processos, de modo mais intenso foi a inserção de tecnologias digitais de comunicação e informação, que permitiram a portabilidade de registros bibliográficos por meio de formatos interoperáveis, como o *Machine Readable Cataloging* (MARC) e a *eXtensible Markup Language* (XML).

Diante dessas inovações, os SOC adquiriram novas perspectivas para o ambiente digital, cuja tarefa não limitava apenas aos processos de descrição e catalogação de materiais. Pois, formatos como o RDF enfatiza a contextualização dos elementos descritos. O RDF é um padrão fundamentado na premissa sujeito, predicado e objeto, que direciona um entendimento para a projeção de relacionamentos semânticos entre conceitos.

Os SOC são agrupamentos com preditivas lógicas e semânticas aplicadas a recuperação da informação. Sob o aspecto tecnológico, os vocabulários controlados tornaram-se a grande

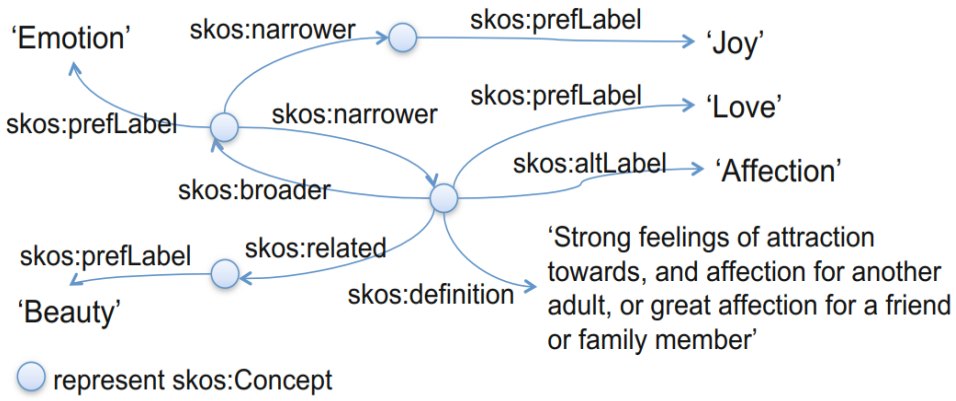
matéria prima para manipulação de sistemas na web por meio de dados referentes a objetos. A utilização de vocabulários interligados ajuda na aquisição de conhecimento por controlar rigorosamente e contextualizar os dados (conceitos, objetos, etc.) (MÉNDEZ; GREENBERG, 2012).

O SKOS que permitem manipular, contextualizar dados e vincular conceitos, inferindo uma ordem semântica, possível de ser interpretada por máquinas de forma simplificada. O que eleva de forma significativa o processo de recuperação e compartilhamento da informação na web. O SKOS fornece um modelo para expressar a estrutura básica e o conteúdo de esquemas de conceito e outros tipos de vocabulários controlados (ISAAC; SUMMERS, 2009).

O SKOS incorpora a característica do RDF, que tornam possível etiquetar termos descritos e por meio de modelos de relacionamento e atribuir um valor semântico entre eles. Segundo Ma et al (2011) para promover funções para indexação e navegação de recursos na Web, seria útil codificar tesauros em formatos compatíveis com a Web. Semelhante ao papel da *Ontology Web Language* (OWL) na edição de ontologias.

O SKOS possui uma característica muito comum com os SOC, pois, com ele é possível criar diferentes tipos de vocabulários controlados (ISAAC E SUMMERS, 2009). Este modelo de representação de informações na web tem como proposta favorecer uma maior reutilização e interoperabilidade entre os vocabulários existentes. (RAMALHO, 2015). Um vocabulário em SKOS contempla alguns aspectos fundamentais, como conceito, termos preferidos e alternativos, termos gerais, específicos, relacionados e notas de definição, conforme apresentada na figura 1, que representa uma visão simplificada de um vocabulário controlado em SKOS.

Figura 1 – Modelo de vocabulário em SKOS



Fonte: Manaf; Bechhofer; Stevens (2012).

As características que qualificam os vocabulários codificados em SKOS precisam de atenção durante o processo de construção, principalmente no que diz respeito ao uso das equivalências semânticas, pois SOCs tradicionais podem conter incongruências que migram para os vocabulários em SKOS. Mastora, Peponakis e Kapidakis (2017) corroboram ao afirmar que a conversão de dados tradicionais para RDF não é uma condição suficiente para adquirir todo o potencial semântico. Ainda mais, em casos como os KOS tradicionais, onde há uma certa falta de semântica bem definida, portanto, torna-se necessário realizar algum tipo de reengenharia na construção ou conversão de vocabulários em SKOS.

3 AVALIAÇÃO PARA VOCABULÁRIO CONTROLADOS

A avaliação de vocabulários controlados é tradicionalmente realizada por especialistas do domínio, no entanto a ANSI/NISO Z39.19 (2010) destaca que há pouca pesquisa ou literatura lidando especificamente com o teste e avaliação de vocabulários controlados. Dentre as formas tradicionais de avaliação de vocabulários controlados encontra-se a própria recomendação da ANSI/NISO Z39.19 (2010), que indica três tipos de avaliação, a saber.

A avaliação heurística, realizada por um perito ou grupo de especialistas, esse é tipo de avaliação mais difundida. O segundo tipo é nomeada modelagem de afinidade, porque reúne uma amostra significativa de usuários, que geram uma classificação analisada contra a hierarquia de termos existentes. O terceiro tipo é o teste de usabilidade, uma avaliação mais ampla, que inclui o sistema de informação utilizado para gerenciar o vocabulário.

Esses tipos de avaliação são bem amplos, não especificando necessariamente adequações necessárias a vocabulários específicos. A ANSI/NISO Z39.19 (2010) cita o trabalho de Soergel (2002) que apresenta uma série de critérios que podem ser adotados no processo de avaliação de um vocabulário controlado. Assim como há outros trabalhos que indicam formas mais precisas de objetivar o processo de avaliação, como Lancaster (2002) que orienta algumas medidas de avaliação que podem ser utilizadas, como razão de equivalência, razão de reciprocidade, definição, flexibilidade, nível de pré-coordenação e tamanho do grupo de termos.

Enquanto Gil-Leiva (2008) e Soler-Monreal (2009) apresentam uma divisão em grupos principais, que denominam de avaliação qualitativa intrínseca e avaliação intrínseca quantitativa ou estatística extrínseca. Para Soler-Monreal; Gil-Leiva (2011) a avaliação intrínseca pode ser realizada com o objetivo de analisar os próprios vocabulários controlados

de modo a estudar sua estrutura, os campos temáticos ou facetas, notas de escopo, relações semânticas, grau de especificidade, etc. Por outro lado, avaliação extrínseca estuda o impacto nos sistemas de informação que os utilizam tanto na indexação quanto na recuperação, (SOLER-MONREAL; GIL-LEIVA, 2011).

Métodos para avaliação tradicional, são plenamente exequíveis nos mais diversos tipos de vocabulários. No entanto, tais técnicas são praticamente inviáveis para as vocabulários modernos, compreendendo milhões de conceitos e relações. (NAYAK et al 2019). Os métodos de avaliação tradicionais possuem uma demanda de tempo e recursos que podem se tornar onerosos, por isso há uma tendência de otimizar cada vez mais a gestão de vocabulários, que inclui sua publicação e manutenção.

De acordo com Nayak et al (2019) há uma necessidade de técnicas automatizadas para avaliar a evolução de hierarquias de conhecimento muito grandes que capturam a subsunção semântica lógica. Quando se aplica ao ambiente digital Pastor-Sanchez; Martinez-Mendez; Rodriguez-Munoz (2012), afirmam que a exploração dos vocabulários na Web é muito limitada. Tais autores ressaltam que o próprio conceito de vocabulário tem evoluído para adaptar-se aos novos modelos de representação da Web, abandonando o paradigma lexical em favor de um paradigma conceitual.

3.1 Parâmetros para avaliação automática de vocabulários em SKOS

O aumento na publicação de vocabulários vinculados no ambiente web revela um grande salto na propagação da interoperabilidade semântica entre SOCs online, entretanto, quando não observadas as regras de manutenção e avaliação, um vocabulário pode se tornar obsoleto, pois trata-se de um produto que opera com linguagem, logo suscetível a variações ao longo do tempo.

A avaliação é uma etapa necessária aos diferentes tipos de SOCs, por isso no âmbito da Ciência da Informação existem distintos tipos de avaliação, que são propostos para atender esta demanda. Incluem-se métodos com abordagens objetivas, subjetivas, automática ou manual. Podendo enfatizar a usabilidade, estrutura, relevância e a qualidade geral de um vocabulário. Quando se trata de vocabulários publicados em SKOS não há métodos consolidados, mas alguns parâmetros que podem ser incorporados no processo de avaliação.

O *SKOS Reference* especifica um número de condições de integridade que devem ser

preenchidas para que o vocabulário seja considerado válido. Muitas dessas condições baseiam-se em padrões anteriores para estruturar vocabulários controlados. (SUOMINEN; HYVÖNEN, 2012). A priori a avaliação de um vocabulário em SKOS pode ocorrer de forma mais generalista, apenas verificando se um vocabulário atende as especificações básicas de uso de um determinado padrão ou esquema de metadados. A seguir são apresentados alguns parâmetros de avaliação encontrados na literatura, que podem auxiliar a análise de um vocabulários em SKOS

Trabalhos como o de Manaf, Bechhofer e Stevens (2012) servem como referência, no que diz respeito a adoção de etapas que podem ser seguidas durante o processo de avaliação de vocabulários em SKOS, pois avaliam os estilos de modelagem usados nesses vocabulários que foram publicados na web. Tal avaliação consiste em preparar um corpus de vocabulário do candidato SKOS, identificar o vocabulários do SKOS, coletar dados de pesquisa, filtrar várias cópias dos mesmos vocabulários do SKOS e analisar o corpus dos vocabulários.

As etapas de avaliação de Manaf, Bechhofer e Stevens (2012) demonstraram, que apesar de existir um grande número de vocabulários publicados, muitos não conseguem se adequar as especificações requeridas em sua estruturação, chamando a atenção de vocabulários vinculados publicados na web que não seguem as recomendações básicas da *World Wide Web Consortium* (W3C), do ponto de vista técnico.

Manaf, Bechhofer e Stevens (2012) e Mastora, Peponakis e Kapidakis (2017) buscaram compreender respectivamente os estilos de publicação de vocabulários e a migração de SOCs tradicionais para o formato SKOS, ambos destacam a necessidade refletir sobre critérios de qualidade que podem ser adotados desde a concepção de um vocabulário, atentando para regras semânticas e cumprimento dos requisitos básicos para uso de determinado padrão.

Quando aplicados aos vocabulários SKOS as avaliações recorrem a verificações estruturais especificadas nas recomendações da W3C para SKOS e as propriedades comuns aos vocabulários controlados. Neste contexto destacam se Mader, Haslhofer e Isaac (2012), Suominen e Mader, (2014) e Suominen e Hyvönen (2012) com abordagens de análises automáticas de vocabulários em SKOS, para isto recorrem a duas ferramentas de análise.

A primeira intitulada *Skosify* utilizada para converter vocabulários em RDFS e OWL para o formato SKOS e melhorar, enriquecer e validar os vocabulários em SKOS. Suominen e Hyvönen (2012) avaliaram 14 vocabulários controlados com esta ferramenta. A segunda chama-se *qSKOS*, utilizada para encontrar problemas de qualidade nos vocabulários do SKOS.

Mader, Haslhofer e Isaac (2012) avaliaram 15 vocabulários com esta ferramenta, enquanto Suominen e Mader (2014) avaliaram 24 vocabulários, todos em SKOS.

Por meio destes respectivos trabalhos é possível apresentar uma compilação de parâmetros aplicáveis aos vocabulários em SKOS. Por que conseguem englobar grande parte dos requisitos essenciais dispostos nas recomendações W3C e da ISO 25964. Como avaliar um vocabulário não é um procedimento orientado por uma ação comum, Suominen e Hyvönen (2012), Mader, Haslhofer e Isaac (2012) e Suominen e Mader (2014) identificam erros recorrentes em vocabulários SKOS que podem ser analisados por meio de verificação automática.

A incidência de tais erros pode comprometer a qualidade de um vocabulário, portanto indica o que não deve conter em um vocabulário controlado. Tais parâmetros endossam o processo de avaliação, neste sentido indica os passos para adoção de parâmetros que visam garantir a qualidade de um vocabulário em SKOS.

Suominen e Hyvönen (2012) identificaram cerca de 14 erros comuns em vocabulários SKOS, categorizados como: URIs válidos, *tags* de idioma ausentes, etiquetas ausentes, conceitos soltos, classes owl disjuntas, uso consistente de etiquetas, uso consistente de propriedades de mapeamento, uso consistente das relações semânticas, valores ambíguos de etiquetas preferidas, sobreposição disjuntiva, propriedades de etiqueta, relações semânticas disjuntivas, hierarquia mais amplas em ciclos e espaço em branco extra (SUOMINEN; HYVÖNEN, 2012).

Mader, Haslhofer e Isaac (2012) classificam os problemas recorrentes em vocabulário SKOS em três categorias mais amplas, especificando os problemas contidos em cada uma delas.

A primeira categoria é classificada como Rotulagem e Documentação, que inclui: *tags* de idioma omitido ou inválido, cobertura incompleta de idiomas, conceitos não documentados e conflitos de etiquetas.

A segunda categoria refere-se a questões estruturais, que incluem: conceitos órfãos, componentes fracamente conectados, relações hierárquicas cíclicas, relações associativas sem valor, conceitos únicos transitivamente relacionados, conceitos superiores omitidos e conceito superior com conceitos mais amplos (MADER; HASLHOFER; ISAAC, 2012).

A terceira aplica-se aos problemas específicos de dados vinculados, e incluem: links

ausentes, links externos ausentes, links quebrados e recursos SKOS indefinidos. Esta categoria refere-se notadamente a problemas de recursos informacionais que podem estar desvinculados dos seus endereços de localização (MADER; HASLHOFER; ISAAC, 2012).

Suominen e Mader (2014) mantém as 3 categorias principais de problemas, e expande sua cobertura, resultando na seguinte classificação:

Etiquetagem e problemas de documentação, abrangendo *tags* de idiomas omitidos ou inválidos, cobertura incompleta de idiomas, conceitos não documentados, etiquetas sobrepostas, etiquetas ausentes, etiquetas preferidas inconsistentes, violação de etiquetas não contíguas e espaços extras vazios em etiquetas (SUOMINEN; MADER, 2014).

Questões estruturais, incluindo conceitos órfãos, clusters de conceito desconectados, relações hierárquicas cíclicas, relações associativas sem valor, conceitos únicos transitivamente relacionados, conceitos superiores omitidos, conceitos principais não marcados, conceitos principais como conceitos mais amplos, conceitos relacionados unidirecionalmente, confrontos relacionais, mapeando confrontos e violação de classes disjuntas (SUOMINEN; MADER, 2014).

E problemas específicos de dados vinculados, como links ausentes, links externos ausentes, links quebrados, recursos SKOS indefinidos, violação do esquema HTTP/URI e URIs inválidos. (SUOMINEN; MADER, 2014).

Outros problemas foram identificados e incluídos no rol de categorias de análise dos problemas de etiquetagem e documentação, nesta instância foram adicionados: caracteres não imprimíveis em etiquetas, etiquetas vazias e referências de notação ambígua. Nas Questões Estruturais, foram incluídas: redundância hierárquica e relações reflexivas.

Para melhor visualizar esse mapeamento de problemas recorrentes, o quadro 1 apresenta uma sistematização categórica dos erros mais comuns identificados em vocabulários em SKOS de acordo com Suominen e Hyvönen (2012), Mader, Haslhofer e Isaac (2012) e Suominen e Mader (2014). Apesar de Suominen; Hyvönen (2012) não categorizarem os erros identificados, estes foram alocados de acordos com suas características comuns.

Quadro 1 – Problemas recorrentes de vocabulários em SKOS

CATEGORIA	PROBLEMA		
	Suominen; Hyvönen (2012)	**Mader; Haslhofer; Isaac (2012)**	**Suominen; Mader, (2014)**
Rotulagem e Problemas de Documentação	*Tags* de idioma ausentes	*Tags* de idioma omitido ou inválido	*Tags* de idiomas omitidos ou inválidos
		Cobertura incompleta de idiomas.	Cobertura incompleta de idiomas
	Propriedades de etiqueta	Conceitos não documentados,	Conceitos não documentados
		Conflitos de etiqueta	Etiquetas sobrepostas
	Etiquetas ausentes		Etiquetas ausentes
	Uso consistente de etiquetas,		Etiquetas preferidas inconsistentes
			Violação de etiquetas não contíguas
	Espaço em branco extra.		Espaços extras vazios em etiquetas
			Caracteres não imprimíveis em etiquetas
			Etiquetas vazias
			Referências de notação ambígua
Questões Estruturais	Conceitos soltos	Conceitos órfãos	Conceitos órfãos
		Componentes fracamente conectados	Clusters de conceito desconectados
	Hierarquia mais amplas em ciclos	Relações hierárquicas cíclicas.	Relações hierárquicas cíclicas
	Uso consistente das relações semânticas	Relações associativas sem valor	Relações associativas sem valor
	Valores ambíguos de etiquetas preferidas	Conceitos únicos transitivamente relacionados,	Conceitos únicos transitivamente relacionados
		Conceitos superiores omitidos	Conceitos superiores omitidos
		Conceito superior como	Conceitos principais não

		conceitos mais amplos	marcados
			Conceitos principais como conceitos mais amplos
	Classes owl disjuntas, sobreposição disjuntiva		Conceitos relacionados unidirecionalmente
			Confrontos relacionais,
	Uso consistente de propriedades de mapeamento		Mapeando confrontos
	Relações semânticas disjuntivas		Violação de classes disjuntas.
			Redundância hierárquica
			Relações reflexivas
Problemas específicos de dados vinculados	URIs válidos	Links ausentes	Links ausentes
		Links externos ausentes	Links externos ausentes
		Links quebrados	Links quebrados,
		Recursos SKOS indefinidos.	Recursos SKOS indefinidos
			Violação do esquema HTTP/URI
			URIs inválidos.

Fonte: elaborado pelos autores.

As questões estruturais envolvem diretamente, problemas de ordem lógica, que comprometem o entendimento contextual dos conceitos em um vocabulário, quando se trata de análises automáticas. É uma das principais preocupações, porque tal falha implica diretamente em erro sintático, semântico e conceitual.

Os problemas específicos a dados vinculados retomam a discussão da manutenção semântica entre os recursos representados, pois a violação ou quebra de relacionamentos prejudica a contextualização do objeto representado. Para Sousa; Martins; Ramalho (2018) os modelos de dados com capacidade semântica constituem-se como ricas fontes de informação além de proporcionar a interoperabilidade na troca dessas informações, portanto, sua manutenção é de grande importância.

A partir dos fundamentos que estruturam os vocabulários controlados, quer sejam normas, diretrizes, padrões ou modelos de representação é possível identificar pontos de referência que sustentarão análises futuras. O resultado dessa pesquisa trata-se de uma fonte referencial de parâmetros que podem ser utilizados para analisar diferentes tipos de vocabulários SKOS. Levando em consideração características fundamentais inerentes aos vocabulários controlados

Diante dos problemas recorrentes na estruturação de um vocabulário que adota SKOS como linguagem codificada de representação, Suominen e Hyvönen (2012), Mader, Haslhofer e Isaac (2012) e Suominen e Mader (2014) fornecem subsídios para identificação dos problemas mais comuns na avaliação de um vocabulário SKOS, que contém as categorias principais dos erros encontrados nos vocabulário em SKOS.

Os principais problemas de etiquetagem para vocabulários SKOS referem-se a inconsistências que podem decorrer do processo de inserção ou conversão dos termos, além de possíveis erros de tradução e informações incompletas inseridas como notas de documentação.

Uma avaliação que consiga detectar esses tipos de erros, evidencia sua importância, principalmente quando aplicadas a vocabulários muito extensos e a validação de conversões para o padrão SKOS. A utilização de métricas automáticas auxiliam na objetivação deste processo, pois simplificam os critérios de avaliação orientados pela própria recomendação que especifica o uso do SKOS.

Os problemas estruturais podem comprometer a semântica de um vocabulário, pois prejudicará a ordem conceitual de um vocabulário com redundâncias lógicas sem valor. Portanto, a identificação desse tipo de erro reduz a probabilidade de inconsistências de organização conceitual.

Os problemas de dados vinculados aplicados aos vocabulários SKOS estão diretamente relacionados com a conectividade entre os recursos informacionais representado e o endereço que indica a sua localização. Logo, trata-se de problemas de links quebrados, ausência de direcionamento e URIs inválidas. A identificação desse tipo de erro, aumenta as chances de recuperação de um recurso informacional e da qualidade dos dados vinculados, que são publicados como vocabulários.

Os vocabulários controlados têm demonstrado grande importância enquanto tipo de

SOC, pois incorporam fundamentos que facilmente se adaptam em novos suportes tecnológicos. O aumento na publicação de vocabulários vinculados no ambiente web revela um grande salto na propagação da interoperabilidade semântica entre SOC. Entretanto, quando não se observa as regras de construção, manutenção e avaliação, um vocabulário pode se tornar problemático do ponto de vista técnico e conceitual.

A qualidade de um vocabulário controlado exige a análise de requisitos específicos dos recursos oferecidos, como infraestrutura, disponibilidade, usabilidade, entre outros fatores, todavia percebe-se uma dificuldade de avaliação de vocabulários de forma holística. Portanto, a apresentação dos parâmetros de avaliação para vocabulários em SKOS destacados neste trabalho oferece um passo que aponta para o avanço dessa discussão.

SKOS é um recurso que potencializa a publicação de vocabulários, amparado por normas internacionais que validam a adoção de modelos simples de organização do conhecimento. Neste sentido, apenas potencializa essa estrutura perene que se encontra no escopo da Organização do Conhecimento. Portanto, este trabalho apresenta de forma resumida esse padrão, com seus principais parâmetros de avaliação.

4 CONSIDERAÇÕES FINAIS

Os vocabulários controlados, enquanto insumos informacionais oriundos da necessidade de resolver problemas de linguagem, que interferem no entendimento de um conceito, porque expressam uma possibilidade de construção sintática e semântica. Essas interferências linguísticas podem ser migradas para o ambiente digital, além da ausência de requisitos que precisam ser atendidos, conforme às especificações exigidas para adoção de um determinado padrão.

As abordagens tradicionais de avaliação de vocabulário são métodos já consolidados, todavia, seu uso em vocabulários com milhares de termos torna-se uma tarefa exaustiva se for realizada manualmente. Tal situação não inviabiliza sua adoção, mas sugere a inclusão de novas métricas de avaliação, como é utilizado no caso dos vocabulários codificados em SKOS. Neste sentido podem ser entendidos como parâmetros que podem ser incorporados em avaliações automáticas ou semiautomáticas.

Os problemas comuns encontrados em nos vocabulários em SKOS são erros que divergem das recomendações para o SKOS, o uso de ferramentas como Skosify e qSKOS

auxiliam nesse o processo de avaliação pois trabalham com uma lógica de análise que verifica se um vocabulário possui determinado erro ou não, caso tenha algum erro, aponta o número de ocorrências encontradas. A incidência de tais erros poderá indicar uma tendência categórica de erros, que pode ser de etiquetagem e problemas de documentação, questões estruturais ou problemas específicos de dados vinculados.

Os procedimentos de avaliação, em etapas sucessivas e utilização de ferramentas automáticas, são parâmetros que podem ser utilizados em vocabulários controlados, principalmente para os que estão codificados em SKOS, tendo em vista, que não se tratam de métodos ou sistematizações utilizadas como instruções normativas, mas produtos de estudos de casos, exploratórios e aplicações técnicas. Por isso, são entendidos como parâmetros que podem ser utilizados como norteadores de processos de avaliação de vocabulários em SKOS.

Os parâmetros apresentados oferecem um ponto de reflexão para considerar os vocabulários controlados atualmente disponíveis na forma dados vinculados (*Linked Data*), levando a questionar os processos de manutenção da qualidade depois de sua publicação, considerando a ênfase que tem sido dado sobre as boas práticas para dados vinculados. Portanto, espera-se que este trabalho possa fomentar as discussões entorno dos SOC e as formas de representação tecnológica disponíveis para a construção de vocabulários controlados.

REFERÊNCIAS

ANSI/NISO. **Z39.19**: guidelines for the construction, format, and management of monolingual controlled vocabularies. Bethesda: NISO, 172, 2010. p. Disponível em: http://www.niso.org/kst/reports/standards?step=2&gid=&project_key=7cc9b583cb5a62e8c15d3099e0bb46bbae9cf38a. Acesso em: 06 mar 2019.

BAKER, Thomas et al. Key choices in the design of Simple Knowledge Organization System (SKOS), Web Semantics: **Science, Services and Agents on the World Wide Web**, v. 20, p. 35-49, 2013. Disponível em: http://www.sciencedirect.com/science/article/pii/S1570826813000176. Acesso em: 19 mar 2019.

GIL-LEIVA, I. **Manual de indización**: teoría y práctica. Gijón, Spain: Trea, 2008.
ISAAC, A.; SUMMERS, Ed. (Ed.). **SKOS Simple Knowledge Organization System Primer**: W3C Working Group Note, 2009. Disponível em: https://www.w3.org/TR/2009/. Acesso em: 19 ago 2012.

LANCASTER, F. W. **Indexação: teoria e prática**. 2. ed. rev. atual. Brasília, DF: Briquet de Lemos, 2004.

LANCASTER, F.W. **El control del vocabulario en la recuperación de la información**. 2. ed. Valencia: Universidad de Valencia. 2002.

MA, Xiaogang et al. A SKOS-based multilingual thesaurus of geological time scale for interoperability of online geological maps. **Computers & Geosciences**, vol. 37, n. 10, 2011. pág. 1602-1615. Disponível em: https://doi.org/10.1016/j.cageo.2011.02.011. Acesso em: 09 abr 2018.

MADER, Christian; HASLHOFER, Bernhard; ISAAc, Antoine. **Finding Quality Issues in SKOS Vocabularies**. TPDL, 2012.

MANAF, Nor Azlinayati Abdul; BECHHOFER, Sean; STEVENS, Robert. **The Current State of SKOS Vocabularies on the Web**. In: E. Simperl et al. (Eds.) ESWC 2012, LNCS 7295, Springer-Verlag Berlin Heidelberg, p. 270–284, 2012.

MASTORA, Anna; PEPONAKIS, Manolis; KAPIDAKIS, Sarantos. SKOS concepts and natural language concepts: An analysis of latent relationships in KOSs. **Journal of Information Science**, v. 43, n.4, p. 492–508, 2017.

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

NAYAK, G., Dutta et al. Automated assessment of knowledge hierarchy evolution: comparing directed acyclic graphs. **Inf Retrieval J**, v.22, 2019. p.256–284. Disponível em: https://doi-org.ez31.periodicos.capes.gov.br/10.1007/s10791-018-9345-y. Acesso em: 6 jul 2019.

NKOS. **Taxonomy of Knowledge Organization Sources/Systems**. 2000. Disponível em: http://nkos.slis.kent.edu/KOS_taxonomy.htm. Acesso em: 6 jul 2019.

PASTOR-SANCHEZ, J. A.; MARTINEZ-MENDEZ, F.J.; RODRIGUEZ-MUNOZ, J.V. Aplicación de SKOS para la interoperabilidad de vocabularios controlados en el entorno de linked open data. **El profesional de la información**, v. 21, n.3, 2012.

RAMALHO, R.A.S. Análise do Modelo de Dados SKOS: Sistema de Organização do Conhecimento Simples para a Web. **Informação & Tecnologia (Itec),** v. 2, p. 66-79, 2015.

SOERGEL, Dagobert. **Thesauri and Ontologies in Digital Libraries**: Tutorial. In: Evaluation of thesauri. Joint Conference on Digital Libraries: Portland, 2002. 107p. Disponível em: http://www.dsoergel.com/cv/B63.pdf. Acesso em: 6 jul. 2019.

SOLER-MONREAL, C; GIL-LEIVA, I. Evaluation of controlled vocabularies by inter-indexer consistency. **Information Research**: An International Electronic Journal, v. 16, n. 4, 2011. Disponível em: http://www.informationr.net/ir/16-4/paper502.html. Acesso em: 10 jan 2019.

SOLER-MONREAL, M. C. **Evaluación de vocabularios controlados en la indización de**

documentos mediante índices de consistencia entre indizadores. Dpto. De Comunicación Audiovisual, Documentación e Historia del Arte Tesis (Doctorado). Universitat Politècnica de València. Valencia, 2009.

SOUSA, J. L.; MARTINS, P. G. M.; RAMALHO, R. A. S. Modelos de representação semântica na era do Big Data. **Brazilian Journal of Information Studies**: ResearchTrends. v. 12, n. 3, p. 34-40, 2018.

SUOMINEN, O., MADER, C. Assessing and improving the quality of SKOS vocabularies. **Journal on Data Semantics**, v. 3, n. 1, p. 47–73, 2014.

SUOMINEN, O.; HYVÖNEN, E. Improving the quality of skos vocabularies with skosify. *In*: International Conference on Knowledge Engineering and Knowledge Management,18.,2012, Ireland. **Proceedings** [...] Ireland: Springer-Verlag, 2012.