# A WEB SEMÂNTICA APLICADA NA RECUPERAÇÃO DE INFORMAÇÃO: Um Estudo de Caso no Contexto Estatístico de Uso de Livros Digitais por Alunos de Graduação

*The Semantic Web Applied in Information Retrieval: A Case Study in the Statistical Context of the Use of Digital Books by Undergraduate Students*

**Stella Schwanz Dias de Assis[1], Alessandra Monteiro Pattuzzo Caetano[2], Henrique Monteiro Cristovão[3]**

(1) Universidade Federal do Espírito Santo - UFES, Vitória/ES, e-mail:stella.assis@edu.ufes.br
(2) Universidade Federal do Espírito Santo - UFES, Vitória/ES, e-mail: apattuzzo@gmail.com
(3) Universidade Federal do Espírito Santo - UFES, Vitória/ES, e-mail: henrique.cristovao@ufes.br

**Resumo:**
Considerando o contexto dos dados estatísticos de uso de livros digitais da biblioteca virtual de uma instituição de ensino superior por seus usuários alunos de graduação, e considerando também o contexto da recuperação de informação na Web semântica em dados ligados RDF, a presente pesquisa tem como objetivo mostrar o processo de mapeamento dos dados existentes, em uma pequena parte da base dessa biblioteca, para dados ligados na Web semântica, dando ênfase ao estabelecimento de interoperabilidade nas camadas sintática, estrutural e semântica. Como prova de conceito, foram executadas consultas sobre a base de dados criada para recuperar informações. Usou-se abordagem qualitativa, natureza aplicada e procedimentos de estudo de caso. Seguindo um *workflow* composto de 10 etapas, desde o conhecimento dos dados e contexto, passando pela elaboração de uma modelagem conceitual simplificada, limpeza, preparação e reconciliação de dados, implementação de uma ontologia operacional, mapeamento RDF, exibição de *kowledge graphs*, até a criação de uma base de dados ligados e a execução de consultas escritas em SPARQL, o percurso da pesquisa conseguiu cumprir o objetivo proposto. Como o desenvolvimento foi realizado em extrato pequeno da base de dados, há necessidade de ampliar a pesquisa, inclusive com a elaboração de uma ontologia de domínio completa para o cenário, bem como a implementação de interface para uso da linguagem de consulta SPARQL, uma vez que ela não é apropriada para o usuário final.

**Palavras-chave:** Recuperação da informação; Web semântica; Interoperabilidade; Acesso à informação; Livros digitais.

**Abstract:**
Considering the context of statistical data on the use of digital books from the virtual library of a higher education institution by its undergraduate students, and also considering the context of information retrieval on the semantic Web in data linked to RDF, this research aims to show the process of mapping existing data, in a small part of the base of this library, to linked data in the semantic Web, emphasizing the establishment of interoperability in the syntactic, structural and semantic layers. As a proof of concept, queries were performed on the database created to retrieve information. A qualitative approach was used with an applied purpose and case study procedures. Following a workflow composed of 10 steps, from knowing the data and context, passing through the elaboration of a simplified conceptual model, cleaning, preparation and reconciliation of data, implementation of an operational ontology, RDF mapping, display of kowledge graphs, until the creation of a linked database and the execution of queries written in SPARQL, the research course was able to fulfill the proposed objective. As the development was carried out in a small extract of the database, there is a need to expand the research, including the elaboration of a complete domain ontology for the scenario, as well as the implementation of the interface for using the SPARQL query language, since that it is not appropriate for the end user.

**Keywords:** Information retrieval; Semantic Web; Interoperability; Access to information; Digital books.

## 1 Introdução

A recuperação de informação (RI) trata da representação, armazenamento, organização e acesso a itens de informação, como documentos, páginas da Web, catálogos online, registros estruturados e semiestruturados, objetos multimídia. A representação e organização dos itens de informação devem ser tais que proporcionem aos usuários fácil acesso às informações de seu interesse (BAEZA-YATES; RIBEIRO-NETO, 2011).

Materializando a Web semântica, os dados ligados surgem como um conjunto de

boas práticas para publicar e conectar conjuntos de dados estruturados na Web, com intuito de criar uma Web de dados ligados (BIZER; HEATH; BERNERS-LEE, 2009).

Nesse contexto de demandas da RI a Web semântica surgiu com o intuito de melhorar a organização da Web e, consequentemente, tornar os seus dados mais facilmente localizáveis inserindo semântica nesses dados com o objetivo que eles sejam melhores compreendidos pelas máquinas (BERNERS-LEE; HENDLER; LASSILA, 2001). Além disso, considera-se que o modelo de dados ligados é unificado e projetados para o compartilhamento global (HEATH; BIZER, 2011) por meio de um planejamento adequado de interoperabilidade das camadas sintática, estrutural e semântica (ZENG, 2019). Interoperabilidade é definido pela National Information Standard Organization (NISO) como "[...] a capacidade de múltiplos sistemas com diferentes hardwares e softwares plataformas, estruturas de dados e interfaces para troca de dados com o mínimo perda de conteúdo e funcionalidade" (NISO, 2004, p. 2, tradução nossa).

A oferta de bibliotecas virtuais vem aumentando no país, e as instituições de ensino superior precisam inovar o processo de acesso e disponibilização da informação. Este é um cenário que não tem volta, e os profissionais bibliotecários, além de disponibilizar, precisam apresentar e acompanhar o uso efetivo dos livros digitais para tomadas de decisão em planejamentos administrativos, financeiros e pedagógicos em suas unidades de informação. Além disso, em um contexto de distribuição de informações na Web, a tarefa de seleção e consulta destas se tornou algo impossível de ser realizado por pessoas, o que torna evidente a necessidade da utilização de aplicações com essa destinação. Nesse sentido, não são estipuladas regras para publicação dos dados, mas apresentadas boas práticas que possibilitam a criação de modelos de dados capazes de se comunicar que sejam interoperáveis. A Web semântica se posiciona como uma extensão da Web (SERRA, 2019).

Em relação ao desenvolvimento de coleções, as decisões de cancelar ou permanecer com o acervo virtual, ocorre durante seu tempo de vigência por meio de controle estatístico de uso e se o acervo disponibilizado é viável para a composição das referências bibliográficas dos planos de ensino das disciplinas dos cursos ofertados pela Instituição de Ensino Superior (IES).

Observa-se uma emergencial necessidade de geração de indicadores que permitam aos profissionais, que atuam em bibliotecas, a elaboração de relatórios e estatísticas de uso dos livros virtuais para a criação de um cenário econômico sustentável pela IES e para dar suporte pedagógico na elaboração dos documentos educacionais essenciais para os funcionamentos dos cursos ofertados pela IES.

Dessa forma, o presente trabalho se propõe a responder ao seguinte problema de pesquisa: como se utilizar dos recursos da Web semântica para organizar os dados disponíveis nas bases de uma biblioteca a fim de fornecer relatórios e estatísticas que auxiliam a gestão da IES ao qual ela pertence?

## 2 Objetivos

Considerando o cenário dos dados e variáveis existentes na base de dados da biblioteca do estudo de caso, a presente pesquisa pretende mostrar, para um pequeno extrato dessa base, o processo de mapeamento dos dados existentes para dados ligados na Web semântica dando ênfase ao estabelecimento de interoperabilidade nas camadas sintática, estrutural e semântica, e uma prova de conceito para a recuperação de informações a partir de uma linguagem de consulta específica de dados ligados.

## 3 Procedimentos Metodológicos

O presente estudo usa abordagem qualitativa, natureza aplicada, sendo utilizado também procedimentos de estudo de caso, que segundo Yin (2001) proporciona um processo investigativo onde são preservadas as características holísticas e significativas dos eventos da vida real.

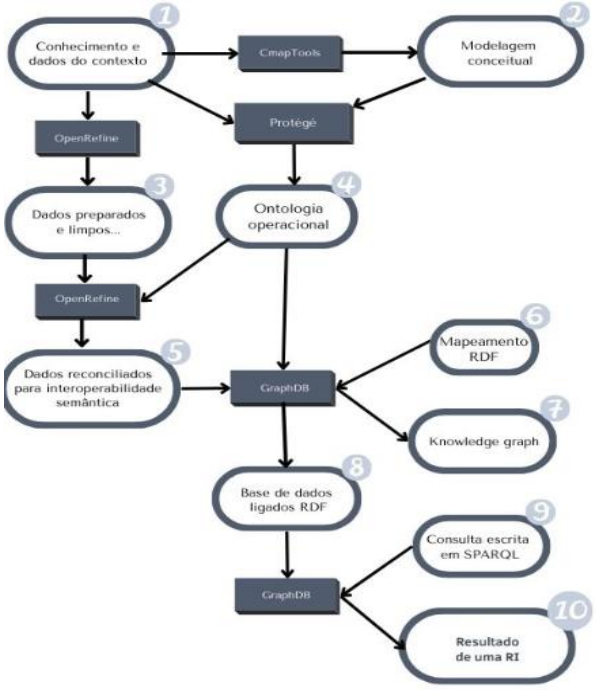
Com o objetivo de testar a proposta desenvolvida utilizou-se uma prova de

conceito pela escrita de consultas de forma a obter um resultado composto de dados de outras bases interligadas por meio de metadados interoperáveis.

O contexto delimitado para a pesquisa está nos relatórios de acesso da biblioteca digital da FAESA - Centro Universitário Espírito Santense, sendo que os dados foram gerados por relatórios da Vital Source[1], provedora do acervo digital, referentes aos acessos do ano de 2021 no primeiro semestre (janeiro a junho).

O *workflow* do desenvolvimento realizado, mostrado na Figura 1, é composto de 10 caixas numeradas que correspondem a etapas, ou elementos informacionais, que alimentam ou que são gerados pelas ferramentas computacionais indicadas pelas caixas escuras, cujas direções de interação estão sinalizadas por setas. Essas 10 etapas do *workflow* são explicados nos próximos parágrafos.

Figura 1 - Workflow do desenvolvimento




Fonte: autoria própria

Etapa 1 - Conhecimento e dados do contexto. Correspondeu ao levantamento dos dados existentes e conhecimento do funcionamento do contexto.

Etapa 2 - Modelagem conceitual. Considerando a redução do cenário dos dados e variáveis existentes na base, foi realizada uma modelagem conceitual simplificada indicando-se as principais entidades do contexto e seus relacionamentos, com o uso da ferramenta CmapTools[2].

Etapa 3 - Dados preparados e limpos. Os dados foram analisados, limpos e preparados com apoio do software OpenRefine[3].

Etapa 4 - Ontologia operacional. A partir da modelagem conceitual e com apoio do software Protégé[4], foi gerada uma ontologia operacional, que segundo Falbo (2014), em sua metodologia para construção de ontologias, denominada de SABiO, corresponde a implementação da ontologia de domínio diretamente em uma linguagem operacional. No caso da presente pesquisa, no contexto dos dados ligados em RDF (Resource Description Framework), foi escolhida a linguagem RDF Turtle[5]. Ainda nessa etapa foram escolhidos elementos interoperáveis da camada sintática e estrutural para incorporação na linguagem. Nessa etapa também buscou-se atender a camada semântica da interoperabilidade pela equivalência de classes.

Etapa 5 - Dados reconciliados para interoperabilidade semântica. Utilizando-se do serviço de reconciliação de dados do OpenRefine, foi executada sobre algumas variáveis escolhidas onde se achou dados abertos correspondentes para acesso. Essa etapa cumpriu, ainda que de forma pequena, a agregação de dados para compor a

[1] VitalSource está disponível em: https://www.vitalsource.com/.

[2] CmapTools é um software para representação de conhecimento. Disponível em: https://cmap.ihmc.us/cmaptools/.

[3] OpenRefine é um software para limpeza, preparação e reconciliação de dados. Disponível em: https://openrefine.org/.

[4] Protégé é um editor de ontologias de código aberto e gratuito. Disponível em: https://protege.stanford.edu/.

[5] RDF Turtle é uma linguagem de marcação para representação de dados ligados RDF. Disponível em: https://www.w3.org/TR/turtle/.

interoperabilidade semântica por via da equivalência de indivíduos.

Etapa 6 - Mapeamento RDF. Nessa etapa usou-se o software GraphDB[6] para inserir um mapeamento RDF com base na ontologia operacional gerada na etapa 4, a fim de gerar um repositório de dados ligados RDF.

Etapa 7 - Knowledge graph. Esse grafo de triplas foi gerado pelo GraphDB, usando dados reconciliados e a ontologia operacional. O intuito é apenas para conferência visual de parte da base, uma vez que é possível visualizar com facilidade as entidades e algumas instâncias com os seus respectivos relacionamentos.

Etapa 8 - Base de dados ligados RDF. Gerada pelo software GraphDB, que se utiliza do mapeamento RDF da etapa 6.

Etapa 9 - Consulta escrita em SPARQL[7]. Como prova de conceito, foram realizadas algumas consultas na base de dados ligados.

Etapa 10 - Resultado de uma RI. É o conjunto de triplas RDF advindas da consulta SPARQL realidade na etapa 9.

## 4 Resultados

Para a modelagem conceitual (etapa 2) selecionou-se as principais entidades do cenário: 'Book', 'Instituição educacional' (como sub categoria de 'Instituição'), 'ISBN', 'Author' (como subcategoria de 'Person'), e 'Publisher'. Em seguida, elas foram implementadas no Protégé, conforme mostra a hierarquia de classes da Figura 2, e incorporadas propriedades necessárias para criar as conexões entre os elementos bem como aspectos de interoperabilidade da camada sintática, a partir de vocabulários amplamente conhecidos como xsd[8], rdf[9], rdfs[10], skos[11], owl[12] entre outros. Também foram incorporados elementos para a camada estrutural de interoperabilidade tais como Dublin Core[13], Friend of Friend (FOAF)[14], Creative Commons(CC)[15], Bibliographic Framework Initiative (Bibframe)[16], Schema.org[17], Data Catalog Vocabulary (DCT)[18] entre outros.

Figura 2 – Hierarquia de classes da ontologia operacional


Fonte: autoria própria, tela capturada do software Protégé

O Quadro 1 mostra algumas equivalências encontradas para adição de interoperabilidade de camada semântica, correspondente à etapa 4 dos procedimentos metodológicos. A estratégia utilizada para a conexão dos dados foi pautada na inserção das propriedades nas anotações dos títulos dos livros, centralizados como principais indivíduos da modelagem.

Quadro 1 - Conexões semânticas entre classes

| Classe | equivalentClass |
|---|---|

[6] GraphDB é um banco de dados orientado a grafo compatível com RDF e SPARQL. Disponível em: https://graphdb.ontotext.com/.

[7] SPARQL é uma uma linguagem de RI estruturada e padronizada para realizar consultas em grafos RDF.

[8] Prefixo e namespace: @prefix xsd: <http://www.w3.org/2001/XMLSchema#>.

[9] Prefixo e namespace: @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.

[10] Prefixo e namespace: @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.

[11] Prefixo e namespace: @prefix skos: <http://www.w3.org/2004/02/skos/core#>.

[12] Prefixo e namespace: @prefix owl: <http://www.w3.org/2002/07/owl#>.

[13] Prefixo e namespace: @prefix dcterms: <http://purl.org/dc/terms/>.

[14] Prefixo e namespace: @prefix foaf: <http://xmlns.com/foaf/0.1/>.

[15] Prefixo e namespace: @prefix cc: <http://creativecommons.org/ns#>.

[16] Prefixo e namespace: @prefix bibframe: <http://id.loc.gov/ontologies/bibframe/>.

[17] Prefixo e namespace: @prefix schema: <http://schema.org/>.

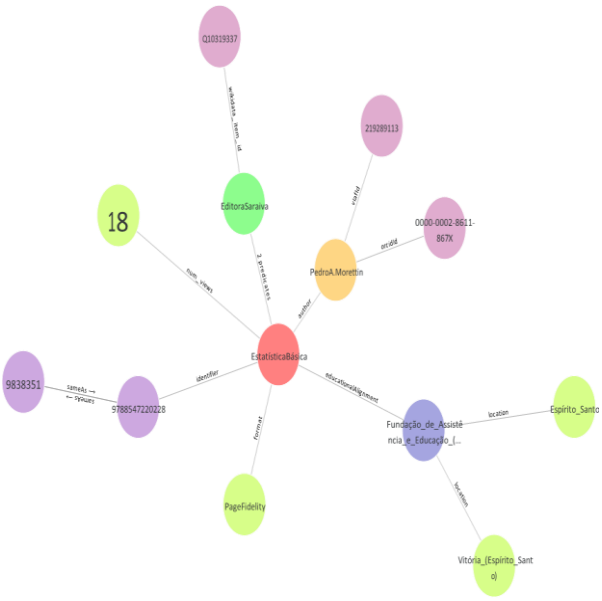
[18] Prefixo e namespace: @prefix dcat: <http://www.w3.org/ns/dcat#>.

| | |
|---|---|
| Author | https://www.wikidata.org/wi ki/Q482980 |
| Book | https://schema.org/Book |
| ISBN | https://id.loc.gov/ontologies /bibframe.html#c_Isbn |
| Instituição Educacional | https://schema.org/Educati onalOrganization |
| Person | https://schema.org/Person %20 http://xmlns.com/foaf/spec/ #term_Person%20 |
| Publisher | http://mappings.dbpedia.or g/server/ontology/pages/O ntologyClass%3APublisher |

Fonte: autoria própria

A fim de agregar uma segunda camada de interoperabilidade semântica, não apenas conectada às classes, mas à cada um dos indivíduos, foi realizada uma reconciliação dos dados, possibilitando uma conexão com diferentes bases, onde foi possível encontrar mais informações sobre elementos reconciliados, desde mais informações sobre os livros através do ISBN, até a busca pelos autores e editoras.

FIGURA 3 - Knowledge graph para conexões de um livro




Fonte: autoria própria, gerado pelo software GraphDB

A Figura 3 mostra um *knowledge graph* para conexões de um livro arbitrário, nesse caso o livro intitulado 'Estatística Básica' do autor 'Pedro A. Morettin'. Esse grafo de triplas RDF corresponde a um possível resultado da etapa 7, e utiliza-se de informações advindas da ontologia operacional da etapa 4, e de dados reconciliados da etapa 5.

Nesse grafo da Figura 3 é possível ver relacionamentos que usam dados reconciliados e, como por exemplo, o nó do indivíduo 'Fundação de Assistência e Educação' conectado por meio da propriedade 'location' com o indivíduo 'Vitória (Espírito Santo)' advindo de uma reconciliação realizada com o software OpenRefine. Outro exemplo é o autor 'Pedro A.Morettin com o relacionamento do seu Orcid, cujo dado também veio de uma reconciliação. Há ainda conexões que estabelecem interoperabilidade semântica entre indivíduos, como é o caso do identificador do livro conectado por meio da propriedade 'sameAs' com o identificador do mesmo livro em outra base de dados.

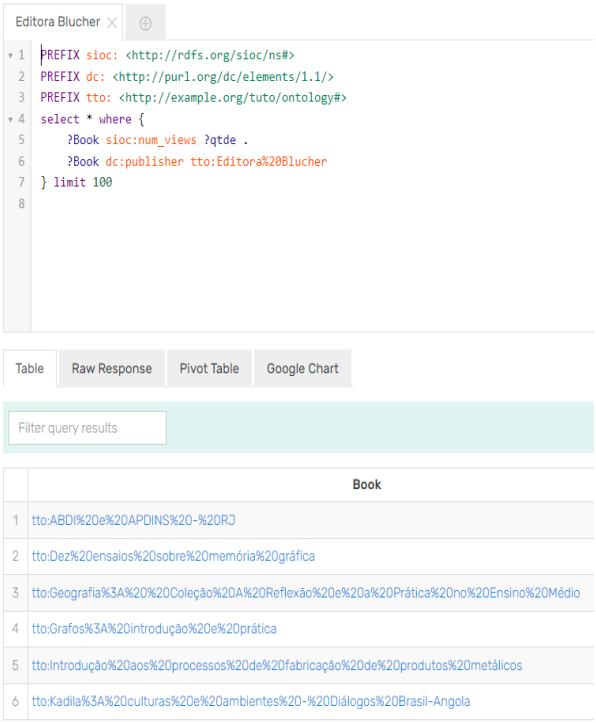
Esse tipo de representação intermediária, por meio de *knowledges graphs*, de pequenos extratos da base de dados ligados, facilita a compreensão e inspeção visual dos relacionamentos entre os dados. Também é possível refletir sobre possibilidades de ampliação de enriquecimento da base de dados ligados com elementos interoperáveis e, assim, possibilitar posteriormente uma RI que agrega elementos de outras bases seja por meio de reconciliação ou interoperabilidade semântica de indivíduos.

A prova de conceito foi obtida realizando-se algumas consultas, escritas em SPARQL, sobre a base de dados ligados no repositório criado no software GraphDB. Uma delas pode ser verificada na Figura 4 inclusive com as triplas obtidas pelo resultado da consulta. Nessa consulta da Figura 4, buscou-se os dados estatísticos da quantidade de empréstimos associados a um determinado livro.

Não obstante, é importante ressaltar também algumas dificuldades encontradas nesse processo. Como foi descrito nos procedimentos metodológicos, foi necessária a realização de uma série de filtragens para obter uma base interoperável. Existe uma carência na alimentação nas bases que possibilitam essa reconciliação, até mesmo

na busca pelas editoras observou-se uma quantidade pequena, além da impossibilidade de uma reconciliação automática de alguns elementos devido a inconsistências nas informações, normalmente por falta de atualização. Outro problema encontrado, que vale destacar, é a falta de acesso à APIs que possibilitam relacionamentos externos, como o exemplo da API fornecido pela OCLC[19] para acessar dados do WorldCat.

Figura 4 - Resultado de uma consulta SPARQL




Fonte: autoria própria, gerado pelo software GraphDB

Resgatando a necessidade informacional que desencadeou essa busca, que seria recuperar informações relevantes, por meio dos dados que levantem possibilidades de estratégias para alcançar um maior número de acessos à base, é possível propor diversas possibilidades com base na exploração e descobertas realizadas.

Uma primeira possibilidade seria reproduzir a interoperabilidade introduzida na base de dados na própria plataforma da biblioteca digital, levando esse recurso ao usuário final. Outra proposta que pode ser agregada é a consolidação de uma base de dados com a bibliográfica básica e complementar de todos os cursos na faculdade, delimitando, além dos cursos, áreas do conhecimento relacionadas e o período em que a obra é exigida. Com uma base que viabilize o acesso à essas informações, é possível enriquecer a base resultante da ontologia, agregando informações que possam contribuir para a construção de algoritmos de recomendação aplicáveis no acervo digital, relacionando os materiais essenciais para os alunos, que podem ser notificados no início de cada período sobre as obras disponíveis, além de possibilitar recomendações baseadas nesses materiais.

## 4 Considerações Finais

A pesquisa possibilitou a verificação positiva da utilização de elementos da Web semântica para inclusão de dados dinâmicos nos catálogos das bibliotecas digitais. Observou-se que as soluções de interoperabilidade evoluíram ao longo do tempo, possibilitando a heterogeneidade de acervos e a preservação da semântica dos conteúdos tornados interoperáveis. Observou-se também que é possível a inclusão de vínculo entre dados de um mesmo catálogo, estabelecendo relações entre registros presentes no acervo e destes com bases de dados externos e de outras ferramentas da Web semântica em bibliotecas, como o Knowledge Graph.

Foram sugeridas propostas de outras possibilidades de enriquecimento na relação interoperável dos dados, evidenciando a necessidade de realização de mais pesquisas sobre a relação entre as bibliotecas digitais e a conexão interoperável de dados da Web.

Quanto ao objetivo proposto, a pesquisa conseguiu mostrar, para um pequeno extrato da base escolhida para o estudo de caso, o processo de mapeamento dos dados

[19] OCLC (Online Computer Library Center, Inc.) é uma organização sem fins lucrativos considerada a maior cooperativa de bibliotecas, museus e arquivos do mundo. Disponível em: https://www.oclc.org/.

existentes para dados ligados na Web semântica dando ênfase ao estabelecimento de interoperabilidade nas camadas sintática, estrutural e semântica. Foi também realizada uma prova de conceito para a recuperação de informações a partir da linguagem SPARQL.

Além disso, é possível inferir que, havendo mais recursos e tempo para continuidade da pesquisa, e utilizando-se o *workflow* apresentado, seria indicado contemplar a base de dados completa bem como elaborar uma ontologia de domínio que considere todo o cenário. Dessa forma, seria possível fazer um acompanhamento estatístico de uso dos livros virtuais disponíveis em uma biblioteca virtual, de forma personalizada para cada objetivo estratégico estabelecido pelo bibliotecário gestor. Seria também necessário implementar uma interface para uso da linguagem de consulta SPARQL uma vez que ela não é apropriada para o usuário final.

**Referências**


BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Recuperação de informação**: conceitos e tecnologia das máquinas de busca. 2. ed. Porto Alegre: Bookman, 2013.

BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim. Linked data - the story so far. **International Journal on Semantic Web and Information Systems** (IJSWIS), v. 5, n. 3, p. 1–22, 2009. Disponível em: https://www.igi-global.com/article/linked-data-story-far/37496. Acesso em: 1 mar. 2021.

FALBO, Ricardo de Almeida. SABiO: Systematic approach for building ontologies. Em: 2014, Rio de Janeiro, RJ. **Anais** [...]. . Em: 1st Joint Workshop ONTO.COM/ODISE on Ontologies in Conceptual Modeling and Information Systems Engineering co-located with 8th International Conference on Formal Ontology in Information Systems. Rio de Janeiro, RJ: CEUR Workshop Proceedings, 2014. p. 14. Disponível em: http://ceur-ws.org/Vol-1301/ontocomodise2014_2.pdf. Acesso em: 13 nov. 2021.

HEATH, Tom; BIZER, Christian. Linked data: evolving the web into a global data space. **Synthesis Lectures on the Semantic Web**: Theory and Technology, v. 1, n. 1, p. 1–136, 2011. Disponível em: https://www.morganclaypool.com/doi/abs/10.2200/S00334ED1V01Y201102WBE001. Acesso em: 2 mar. 2021.

NISO. **Understanding metadata**. Bethesda (EUA): National Information Standards Organization, 2004. Disponível em: http://www.niso.org/publications/press/UnderstandingMetadata.pdf . Acesso em: 13 nov. 2021.

SERRA, Liliana Giusti. A web semântica na gestão de livros digitais licenciados: uma proposta de modelo. 2019. Tese (Doutorado em Ciência da Informação), Programa de Pós-Graduação em Ciência da Informação da UNESP, São Paulo, 2019. Disponível em:https://repositorio.unesp.br/handle/11449/183526. Acesso em: 02 nov. 2021.

YIN, Robert K. **Estudo de caso**: planejamento e métodos. 3. ed. Porto Alegre, RS: Bookman, 2001

ZENG, Marcia Lei. Interoperability. **Knowledge Organization**, v. 46, n. 2, p. 122–146, 2019. Disponível em: http://www.isko.org/cyclo/interoperability. Acesso em: 13 nov. 2021.