DISPONIBILIZAÇÃO DE METADADOS EM LINKED DATA PARA REPOSITÓRIOS DIGITAIS

Felipe Augusto Arakaki*

Faculdade de Ciência da Informação, Campus Universitário Darcy Ribeiro, Universidade de Brasília.

Caio Saraiva Coneglian**

Faculdade de Filosofia e Ciências, Universidade Estadual Paulista.

Plácida Leopoldina Ventura Amorim da Costa Santos***

Faculdade de Filosofia e Ciências, Universidade Estadual Paulista.

José Eduardo Santarem Segundo****

Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo.

Resumo: A considerar a expansão da produção científica em ambientes informacionais digitais e as novas formas de disponibilização de dados seguindo os princípios do Linked Data, objetiva-se discutir as possibilidades de relacionamentos de datasets e o enriquecimento semântico de metadados em repositórios digitais. Ainda, expor um modelo de conversão de registros em RDF. É uma pesquisa teórica e exploratória, a qual realiza uma revisão bibliográfica sobre repositórios digitais e Linked Data. Desta forma, demonstraram-se possibilidades do processo de conversão, com a identificação de bases de dados, vocabulários e padrões que devem ser adotados para que os dados gerados sejam enriquecidos semanticamente. O trabalho apresentou um modelo que reflete os passos que devem ser aplicados no processo de disponibilização dos metadados de um repositório digital em Linked Data. Conclui-se que a integração entre os repositórios digitais e as tecnologias da Web Semântica permite a disponibilização de dados em Linked Data, que fornece novos meios para a divulgação e a integração dos recursos na Web.
Palavras-chave: Web Semântica, Linked Data, repositórios digitais, metadados.

Título: DISPONIBILIDAD DE METADATOS EN *LINKED DATA* PARA REPOSITORIOS DIGITALES.
Resumen: Considerando la expansión de la producción científica en ambientes informacionales digitales y las nuevas formas de disponibilización de datos siguiendo los principios *Linked Data*, se objetivó discutir posibilidades de relacionamiento de *datasets* y enriquecimiento semántico de metadatos en repositorios digitales. Adicionalmente, presentar un modelo de conversión de registros en RDF. Es una investigación teórica y exploratoria, que realizó una revisión bibliográfica sobre repositorios digitales y *Linked Data*. Se demostraron posibilidades del proceso de conversión, con la identificación de bases de datos, vocabularios y estándares que deben ser adoptados para que los datos generados sean enriquecidos semánticamente. El trabajo presenta un modelo que refleja los pasos a ser aplicados durante el proceso de disponibilización de metadatos de un repositorio digital en *Linked Data*. Se concluye que la integración entre repositorios digitales y las tecnologías de la web semántica permite la disponibilización de datos en *Linked Data*, la cual proporciona nuevos medios para la divulgación y la integración de los recursos en la web.
Palabras clave: Web Semántica; *Linked Data*; repositorios digitales; metadatos.

Title: AVAILABILITY OF METADATA IN LINKED DATE FOR DIGITAL REPOSITORIES.
Abstract: Considering the expansion of scientific production in digital information environments and the new forms of data availability following the principles of Linked Data, the objective is to discuss possibilities of relationships of datasets and semantic enrichment of metadata in digital repositories and present a model of conversion of records in RDF. It is a theoretical and exploratory research, performing a bibliographic review on digital repositories and Linked Data. In this way, we have demonstrated possibilities of the conversion process, data is enriched with the identification of databases, vocabularies and standards that must be adopted for the data generated to be semantically enriched. The work presented a model that reflects the steps to be applied in the process of making available the metadata of a digital repository in Linked Data. It is concluded that the integration between digital repositories and Semantic Web technologies allows the availability of data in Linked Data, which provides new means for the dissemination and integration of resources on Web.
Keywords: Semantic Web; Linked Data; digital repositories; metadata.

* fe.arakaki@gmail.com
** caio.coneglian@gmail.com
*** placidasantos@gmail.com
**** santarem@usp.br

Recibido: 10-04-2018; 2ª versión: 03-07-2019; aceptado: 15-07-2019.

ARAKAKI, F.A.; CONEGLIAN, C.S.; SANTOS, P.L.V.A. y SANTAREM SEGUNDO, J.E. Disponibilización de metadatos en *Linked Data* para repositórios digitales. *Anales de Documentación*, 2019, vol. 22, nº 2. Disponible en: http://dx.doi.org/10.6018/analesdoc.327441.

1 INTRODUÇÃO

A produção científica tem crescido gradativamente nas últimas décadas, em especial pelo aumento dos cursos de pós-graduação, nos números de revistas científicas e na quantidade de eventos que ocorrem em todo mundo. As instituições de ensino começaram, assim, a empenhar-se em criar mecanismos para reunir e divulgar a sua produção intelectual na Web, a fim de ampliar sua visibilidade.

Neste cenário, os repositórios digitais foram desenvolvidos com o intuito de auxiliar neste processo, ao facilitar o armazenamento e a disponibilização dos conteúdos em acesso aberto. Outro elemento primário são os metadados usados para descrever os recursos armazenados. Em suma, os repositórios digitais utilizam padrões de metadados com características próprias da Web, como sintaxe XML e tags textuais. O principal exemplo é o padrão Dublin Core, aplicado pela maioria dos repositórios digitais para representarem os seus recursos e promover a interoperabilidade com outros ambientes informacionais digitais.

No âmbito da interoperabilidade, o uso de padrões de metadados possibilita que diversos sistemas possam compreender a descrição de um recurso de forma semelhante. Desta maneira, permite que informações sejam trocadas com mais facilidade. Ademais, assegura a disponibilização dos recursos informacionais seguindo os princípios do Linked Data.

O Linked Data pode ser considerado como princípios de como realizar ligações entre recursos. Utilizam-se os metadados e as tecnologias da Web Semântica para possibilitar semântica e para que possam ser compreensíveis pelos computadores. Nos ambientes digitais, os conjuntos de metadados contemplam informações diversificadas sobre o recurso informacional, facilitando o relacionamento a outras bases de dados a partir de informações, como autores e assuntos. Além disso, a utilização dos conceitos do Linked Data em repositórios é capaz de expandir e explicitar as relações existentes entre bases de dados.

Com a estruturação de padrões de metadados aderentes aos princípios do Linked Data, além do enriquecimento com ontologias, seria possível que a localização e a integração de bases distintas ocorram naturalmente. Promove, entre outros benefícios, a possibilidade de realizar inferências e serendipidade. Neste contexto, questiona-se: como os repositórios digitais podem estruturar seus metadados no contexto das boas práticas do Linked Data, visando o enriquecimento semântico do dataset?

2 OBJETIVOS E METODOLOGIA

2.1 OBJETIVOS

Desta forma, este trabalho tem como objetivo discutir as possibilidades de relacionamentos de datasets e o enriquecimento semântico de metadados em repositórios digitais, no contexto do Linked Data por meio das tecnologias da Web Semântica. Nesse cenário, após as discussões sobre o uso do Linked Data em repositórios e enriquecimento semântico, é apresentado um modelo para conversão dos registros do repositório em uma estrutura em RDF. Oferece, assim, uso efetivo dos dados convertidos em Linked Data.

2.2 PROCEDIMENTOS METODOLÓGICOS

Como procedimentos metodológicos, caracteriza-se como uma pesquisa teórica e exploratória, a partir de um levantamento bibliográfico sobre a aplicação das tecnologias do Linked Data em repositórios digitais. Dessa forma, foram utilizadas fontes bibliográficas localizadas por meio das bases de dados: base de dados em ciência da informação (BRAPCI), Scopus, Web of Science, Portal de Periódicos CAPES, no período de 2006 (início do Linked Data) até 2017. O levantamento bibliográfico permitiu realizar um apanhado sobre os conceitos envolvidos (repositórios digitais, Linked Data) e como se relacionam.

Após as discussões e os resultados é apresentado um modelo para conversão de registros de um repositório digital para uma estrutura em RDF. Isso possibilitará que outras bases possam consumir os dados do repositório em Linked Data.

3 WEB SEMÂNTICA E LINKED DATA

A Web Semântica surgiu a partir da proposta de Berners-Lee, Hendler e Lassila, em 2001. Identifica-se como uma extensão da Web e visa aprimorar a forma como os mecanismos computacionais compreendiam os conteúdos construídos por pessoas. Torna, assim, a Web em um ambiente mais interoperável. (Berners-Lee; Hendler; Lassila, 2001). Desde a sua concepção, uma série de tecnologias foi desenvolvida na tentativa de ser implementável a proposta de Berners-Lee, Hendler e Lassila (2001). Tais tecnologias abarcavam desde linguagens para a estruturação e a representação dos conteúdos, como a eXtensible Markup Language (XML) e o Resource Description Framework (RDF), passando por protocolos para recuperação de dados, como o SPARQL Protocol and RDF Query Language (SPARQL) e as linguagens para a construção de ontologias, como o Web Ontology Language (OWL). Chega até estruturas básicas para a identificação única de conteúdos, como o Uniform Resource Identifier (URI).

As tecnologias e ferramentas apresentadas foram desenvolvidas e amadurecidas com grupos de trabalhos, principalmente do World Wide Web Consortium (W3C). Criaram-se conjuntos de especificações, que promoveram a disseminação nas comunidades de desenvolvedores. Desta forma, essas tecnologias passaram a ser utilizadas em diversas aplicações, que buscavam materializar de alguma forma a Web Semântica. (Santarem Segundo, Coneglian, 2016).

A iniciativa que melhor representa a materialização da Web Semântica é o Linked Data. A proposta do Linked Data nasceu em 2006 por Tim Berners-Lee, numa discussão sob a finalidade da Web Semântica de realizar ligações expressivas entre os conteúdos da Web. Propôs-se o Linked Data como forma de ligar os conteúdos, a considerar que no futuro os usuários fossem capazes de encontrar informações relacionadas. Arakaki (2016, p. 27) afirma que o Linked Data: "[...] diz respeito em como ligar dados. Configura, assim, como melhores práticas para estruturar e ligar dados. Esse processo facilita a busca de agentes humanos e não humanos e os direcionam em diferentes bases a partir desses dados ligados."

Berners-Lee (2006) ainda apresentou os princípios do Linked Data neste texto:

1. *Use URIs como nomes para as coisas;*
2. *Use HTTP URIs para que as pessoas possam procurar esses nomes;*
3. *Quando alguém procurar um URI, fornecer informações úteis, usando os padrões (RDF, SPARQL);*
4. *Incluir links para outros URIs, para que eles possam descobrir mais coisas.*

Estes quatro princípios revelam as principais tecnologias que envolvem o *Linked Data*. Destacam-se os URIs para identificar e ligar os recursos e os padrões RDF e SPARQL para representar e promover uma recuperação, seguindo a ideia dos conceitos da Web Semântica. Tal explanação indica que o Linked Data está focado essencialmente na ligação de dados, a partir de determinados princípios e boas práticas. A Web Semântica está presente principalmente ao fornecer tecnologias para o estabelecimento desta proposta, ao passo que influencia ao fornecer ao Linked Data os seus princípios quanto ao estabelecimento da chamada Web de Dados.

No entanto, a disponibilização de conjuntos de dados em Linked Data exige uma série de processos para que os dados publicados estejam enquadrados como Linked Data e com um significado explícito nas relações. O principal processo neste contexto é o enriquecimento semântico.

O enriquecimento semântico visa expandir as relações do conjunto de dados e normalmente ocorre após a conversão dos dados em RDF. Neste sentido, os dados em RDF não contêm uma carga de ligações expressiva com outras bases de dados, nem permitem a realização de inferências e axiomas. O enriquecimento deverá, assim, possibilitar a realização de ligações com outros conjuntos de dados. Ainda, permitem às propriedades semânticas, principalmente de ontologias, a inserção de axiomas nos dados, para oferecer uma futura realização de inferências.

Lóscio, Burle e Calegari (2017) complementam que há uma gama ampla de técnicas de enriquecimento, ao dizer que estas são complexas. Os autores destacam ainda técnicas de aprendizagem de máquinas tenham sido utilizadas, com o princípio de melhorar o enriquecimento.

Atualmente, há uma série de ferramentas que auxiliam no processo de publicação de dados no âmbito do Linked Data. Um projeto denominado LOD2 reúne um conjunto de tecnologias que podem auxiliar nesta tarefa. Apresenta ferramentas que vai desde o processo de extração, passa pelo enriquecimento e chega até a publicação dos dados. (LOD2, 2017).

Além disso, outros pontos também devem ser seguidos para que a publicação dos dados na Web siga os princípios do Linked Data e da Web Semântica.

Para isso, em 2014 foi publicado pela W3C as "boas práticas para publicação do linke data", com o intuito de estruturar passos para publicação de dados em Linked Data em um contexto geral para Web. (Hyland; Atemezing; Villazón-Terrazas, 2014). Outro documento desenvolvido sob a égide da W3C apresenta as chamadas "boas práticas para a publicação de dados na Web". Demonstra, desta forma, quais pontos devem ser considerados ao serem publicados datasets. (Lóscio; Burle; Calegari, 2017).

Neste documento, os autores apresentam 35 boas práticas divididas em 12 categorias, que vão desde questões de representação dos dados, passa por enriquecimento semântico e chega até a preservação e acesso dos datasets. As boas práticas afirmam que ao seguir as boas práticas, os conjuntos de dados terão oito benefícios (compreensão, processabilidade, descoberta, reuso, veracidade, capacidade de ligação, acesso e interoperabilidade).

O desenvolvimento destas boas práticas é fundamental para que os conjuntos de dados disponibilizados na Web sejam cada vez mais estáveis. Assim, podem difundir a publicação de dados conforme os princípios do Linked Data. A publicação de dados pode ocorrer em diversos domínios e cenários, como nos dados médicos e nos dados bibliográficos.

No que tange ao domínio bibliográfico, diversas pesquisas investigam criar meios de integrar os registros nas tecnologias da Web Semântica e do Linked Data. Neste contexto, Baker et al. (2011) afirmam que:

> *A web de informações deve ser adotada, tanto disponibilizando dados disponíveis como dados vinculados e usando a web de dados em serviços de informação. Idealmente, os dados devem integrar-se integralmente com outros recursos na Web [...] Ao se envolver com a Web de Linked Data, as bibliotecas podem assumir um papel de liderança fundamentado em suas atividades tradicionais: gerenciamento de recursos para uso atual e preservação de longo prazo; descrição dos recursos com base em regras acordadas; e respondendo às necessidades dos buscadores de informações.*

Os autores demonstram processos inerentes aos serviços de informação e às bibliotecas que podem estar integrados ao Linked Data, como questões de descrição e de representação das informações. Neste sentido, iniciativas como o BIBFRAME e o Dublin Core RDF contribuem para implementar esta integração. Fornece padrões de metadados e vocabulários que permitem a representação de recursos bibliográficos, de acordo com as necessidades deste domínio.

No contexto do domínio bibliográfico, os repositórios digitais recebem destaque e são difundidos como um meio eficiente de disponibilizar os dados integrados com as tecnologias atuais, em especial a Web. Um repositório digital é um ambiente propício para integrar os princípios da Web Semântica, para que assim possa ocorrer a publicação de datasets com os metadados dos registros existentes. Visa aproximar essa relação entre os repositórios digitais e o Linked Data, e na sequência apresentam-se conceitos sobre os repositórios digitais.

4 REPOSITÓRIOS DIGITAIS

Com o movimento de acesso aberto, houve uma expansão e popularização dos repositórios digitais para disponibilizarem trabalhos acadêmicos, técnico-científicos, entre outros tipos de materiais, principalmente no contexto de universidades e centros de pesquisa. "A expressão 'repositórios digitais' [...] é empregada para denominar os vários tipos de aplicações de provedores de dados que são destinados ao gerenciamento de informação científica, constituindo-se, necessariamente, em vias alternativas de comunicação científica." (Leite, 2009, p. 19). Segundo o CONARQ (2015, p. 9)

> *[...] repositório digital é um ambiente de armazenamento e gerenciamento de materiais digitais. Esse ambiente constitui-se de uma solução informatizada em que os materiais são capturados, armazenados, preservados e acessados. Um repositório digital é, então, um complexo que apoia o gerenciamento dos materiais digitais, pelo tempo que for necessário, e é formado por elementos de hardware, software e metadados, bem como por uma infraestrutura, organizacional e procedimentos normativos e técnicos.*

Gonçalez (2017, p. 79-82) complementa que os repositórios digitais

> *[...] são formados a partir de coleções digitais que podem ser construídas de formas e com propósitos diferentes, porém atuam como provedores dos dados que fazem o gerenciamento da informação. [...] O repositório digital favorece o armazenamento de um número grande de documentos, possibilitando assim que recursos sejam disponibilizados e preservados de forma que esses materiais possam ser acessados em longo prazo. [...] Diante do exposto, temos a clareza de que os repositórios digitais são sistemas de informação que*

fazem o gerenciamento e o armazenamento de coleções de objetos digitais, por um longo período de tempo e proveem o acesso apropriado.

Para Leite (2009), os repositórios digitais podem ser divididos em Repositórios Temáticos ou Disciplinares, que reúnem conjunto de objetos digitais de determinadas áreas dos conhecimentos; Repositórios de Teses e de Dissertações, que compõe as teses e dissertações defendidas em Programas de Pós-Graduação; e em Repositórios Institucionais, que concentram a produção de uma instituição. Entretanto, com a perspectiva de disponibilização de dados científicos, estes repositórios começaram a ter grande destaque.

Entre as principais características dos repositórios estão a disponibilização de conteúdos em acesso aberto e a realização do autoarquivamento. Todavia, algumas instituições, como a Universidade Estadual Paulista "Júlio de Mesquita Filho" (Unesp), têm adotado outras metodologias para alimentação de seu repositório, a partir de coletas automáticas, conforme apontado por Vidotti et al. (2016). Assim como o uso de APIs para importação de dados de outras bases de dados.

Nesse contexto, os padrões metadados em repositórios são essenciais, pois facilitam a interoperabilidade entre sistemas. Diversos autores já discutiram a importância e as funções dos metadados, como Baca (1998, 2016), Duff e McKemmish (2000), Mendez Rodrigues (2002), Duval et al. (2002), Chowdhury, G. e Chowdhury, S. (2007), Zeng e Qin (2008; 2016), Alves (2005; 2010), Alves e Santos (2013), Simionato (2012; 2015), Pomerantz (2015) e Riley (2017), entre outros.

Dessa forma, entre as principais funções dos metadados destacadas foram: descrever e facilitar a identificação de um recurso e a autenticação dos registros; garantir a persistência de registros de conteúdo, estrutura e contexto; possibilitar a gestão dos termos e condições de acesso e eliminação de recursos informacionais; acompanhar a documentação da história da utilização dos recursos informacionais, incluindo os processos de registros e arquivamento; auxiliar os usuários na descoberta e recuperação de recursos informacionais; restringir o uso não autorizado; auxiliar na preservação da informação; e facilitar a interoperabilidade em ambientes informacionais.

Dentre os primeiros padrões para descrição de recursos digitais está o Dublin Core. Sua estrutura flexível, modular e o consenso internacional apontados por Weibel (1995) e Baptista e Machado (2001) proporcionam a utilização dos princípios do Linked Data, conforme destacado por Baker (2012).

Essas características do Dublin Core possibilitam a interoperabilidade entre sistemas. Segundo Miller (2000), a interoperabilidade consiste no processo em que os sistemas, os procedimentos e a cultura de uma organização sejam gerenciados de forma a maximizar as oportunidades de intercâmbio e a reutilização de informação, seja interna ou externamente. Há diversos tipos de interoperabilidade, como o que aborda questões de padrões de comunicação, transporte, armazenamento e representação como o protocolo Z39.50; interoperabilidade semântica, que levanta questões no uso dos metadados para descrever conceitos similares como autor, criador, compositor; interoperabilidade política/humana que abrange como a informação é descrita e divulgada; interoperabilidade intercomunitária que aborda questões de interoperabilidade entre comunidades; interoperabilidade legal que está relacionada aos requisitos legais de distribuição e licença dos conteúdos; e, por fim, a interoperabilidade internacional que traz questões de troca de informações de nível internacional. Baker et al. (2011) relatam ainda que enquanto Linked Data refere-se à interoperabilidade técnica dos dados, o Linked Open Data centra na interoperabilidade legal.

Segundo a Confederation of Open Access Repositories (2017), há vários benefícios para que repositórios publiquem seus dados na proposta do Linked Open Data. Entre eles estão: desenvolver serviços locais e mais amplos em recursos, agregando informações; integrar diferentes tipos de informações, como os recursos bibliográficos, as estatísticas ou as informações geoespaciais; enriquecimento de dados de outras fontes de dados vinculados, especialmente vocabulários controlados, dados de autoridade e padrões de codificação de sintaxe; aumento da recuperação do repositório institucional pelos motores de busca na Web; as coleções são mais fáceis de acessar e, ao mesmo tempo, tornam mais úteis as novas aplicações; redução da redundância de descrições bibliográficas na Web.

Para Byrne e Goddard (2010), vincular dados a ontologias que descrevem propriedades e relacionamentos possibilita que os computadores entendam o conteúdo e ainda obtenham novos conhecimentos sobre. Por exemplo, Monteiro Lobato escreveu "Sítio do Picapau Amarelo", logo o inverso também deve ser inferido, ou seja, o livro "Sítio do Picapau Amarelo" foi escrito por Monteiro Lobato.

Um caso foi relatado por Latif, Borst e Tochtermann (2014) de aplicação do Linked Data em um repositório de acesso aberto da área da economia, o EconStor, gerenciado pela Biblioteca Nacional da Alemanha. Segundo os

autores, há três formas diferentes de converter dados vinculados de um banco de dados relacional (estrutura base do DSpace). Em dois casos é obrigatório um arquivo de mapeamento dos dados do banco de dados relacional para formação de triplas em RDF, como o uso das plataformas D2RQ ou Open Link Virtuoso que fazem uma transformação síncrona de um banco de dados relacional para RDF, ou seja, é feita uma transformação em tempo real das consultas sobre o resultado. Outra forma de conversão do banco de dados relacional para o modelo RDF é por meio do Triplify, que funciona com instantâneos e visualizações sobre os conteúdos do banco de dados relacionais que são expostos como gráficos RDF. A terceira categoria funciona nativamente em triplas da *Web* semântica, e os dados podem ser migrados de um banco de dados relacional usando o R2RML. (Latif; Borst; Tochtermann, 2014).

Diante das questões apresentadas, a seguir demonstram-se os resultados e as discussões do presente trabalho, em que é explicada a aplicação dos princípios do Linked Data em Repositórios Digitais.

5 RESULTADOS E DISCUSSÕES: APLICAÇÃO DOS PRINCÍPIOS DO *LINKED DATA* EM REPOSITÓRIOS DIGITAIS

Transpor os recursos bibliográficos e arquivísticos para os formatos da Web Semântica, em especial do Linked Data, exige reflexões e a definição de estruturas dos dados, de forma a trazer os benefícios dessas propostas aos conjuntos de recursos. Desta maneira, o processo de conversão para o Linked Data extrapola uma simples migração de formato. Passa, necessariamente, pela identificação de bases de dados que podem ser vinculadas aos recursos informacionais, à definição de uma estrutura dos dados que segue os princípios e às boas práticas para a publicação de dados. Além de ter um processo de enriquecimento, em que se busca aumentar o nível de semântica formal dos dados.

Neste sentido, os dados e metadados dos repositórios digitais são importantes fontes de informações que podem ser transpostas para o âmbito do Linked Data, pela visibilidade que estes instrumentos fornecem às produções intelectuais de uma instituição ou de uma temática específica. A conversão desses registros pode aumentar a visibilidade dos recursos, ao mesmo tempo em que permite que novos conhecimentos sejam gerados a partir de inferências e associações entre os dados. Vale destacar que os repositórios digitais possuem particularidades, que devem ser consideradas e tratadas quando se busca migrar os seus registros para o Linked Data.

Com o intuito de demonstrar como pode ocorrer esse processo no contexto dos repositórios digitais, esta seção é dividida em duas partes. A primeira discute as principais fontes de informação que podem ser vinculadas aos registros, e como devem ser estruturados os dados seguindo as particularidades do domínio em questão. Em seguida, apresenta-se uma proposta sobre os passos que devem ser adotados para a conversão dos registros.

5.1 Definição de datasets e Estruturação dos Dados

Para a construção de um repositório, a estruturação a partir de metadados é fundamental para a localização, a recuperação, entre outros aspectos, conforme apontado anteriormente. Portanto, o uso de um perfil de aplicação é fundamental para minimizar os problemas de interoperabilidade entre sistemas. Segundo Coyle e Baker (2009), um perfil de aplicação aborda documentos que vão desde o levantamento dos requisitos funcionais, passando pela construção de um modelo de domínio que mapeará as entidades descritas e como se relacionam. Posteriormente, enumera os termos de metadados a serem empregados e as regras para a sua utilização (descrição conjunto de perfis e diretrizes de uso). Ademais, define a sintaxe que será utilizada para codificar os dados (diretrizes de sintaxe e formatos de dados). Nesse âmbito, o Modelo Abstrato Dublin Core (Dublin Core Abstract Model - DCAM) que prevê a estruturação dos dados em RDF na camada dos padrões de fundamentação, influencia diretamente os vocabulários de metadados nos perfis de aplicação.

Outro ponto necessário para a conversão dos registros para os princípios do Linked Data está na ligação dos registros com outras bases de dados e datasets. Isto é fundamental para o estabelecimento do Linked Data, uma vez que os relacionamentos entre recursos permitem que os usuários naveguem e obtenham informações detalhadas de cada informação que está relacionada.

Neste sentido, há bases de diversos contextos que podem ser relacionadas aos elementos, como bases para autores, para identificadores e para assuntos. Destacam-se as bases de autores, em que algumas delas são construídas a partir dos princípios do Linked Data, como o Virtual International Authority File (VIAF) e o Open Researcher and Contributor ID (ORCID).

No entanto, o uso do VIAF em repositório pode ser dificultado por não conter estudantes de pós-graduação e pesquisadores, pois essa base está focada em autores de livros. Além disso, é difícil incluir um novo autor nessa base,

visto que há um grupo restrito de bibliotecas nacionais que têm essa autorização. Em paralelo ao VIAF está o International Standard Name Identifier (ISNI), um número padrão certificado pela ISO 27729 para identificar nomes de pessoas, como pesquisadores, inventores, escritores, artistas, criadores visuais, intérpretes, produtores, editores, agregadores e outros. (Arakaki, 2016).

Uma das alternativas para a identificação de autores pode ser a utilização de identificadores de currículo nacionais como o Lattes (http://lattes.cnpq.br/) no Brasil e o CIÊNCIA VITAE (https://cienciavitae.pt/) em Portugal. São fontes ricas para realizar ligações, para que os usuários possam navegar ao perfil de um pesquisador. Por outro lado, a ORCID (https://orcid.org/) é uma base de autores de âmbito internacional que permite realizar ligações com registros.

Com relação às bases para identificar as produções científicas, o Digital Object Identifier (DOI) é um identificador único de recursos digitais desenvolvido e administrado pela CrossRef. Permite que os recursos não sejam confundidos com outros ou se tornem ambíguos na sua identificação. Similar ao DOI, o Handle System é um sistema distribuído concebido para assinalar, armazenar, administrar e resolver problemas de identificadores e garantir nomes persistentes a objetos digitais. (SAYÃO, 2007).

Em relação a assuntos e sistemas de classificação, a depender da área, pode-se utilizar tesauros e vocabulários controlados na proposta do Linked Data como Library of Congress Subject Headings (LCSH), AGROVOC, Classificação Decimal Universal (CDU), entre outros.

As bases de dados apresentadas são elementos essenciais quando se pretende estruturar os registros de repositórios digitais nos princípios do Linked Data, pois elas vão inserir esses registros integrados a alguns dos principais projetos de Linked Data existentes a nível global. Além disso, essas ligações possibilitam que, ao navegar pelos dados, os usuários possam encontrar informações mais específicas e com alto nível de confiabilidade.

Em suma, o Quadro I apresenta algumas das principais bases de dados que podem ser utilizadas no processo de conversão dos registros de repositórios digitais para os Linked Data.

Tipo de Informação	**Bases de Dados**
Bases de autores e instituições	VIAF, ORCID, Lattes, ISNI
Identificadores produções científicas	DOI, Handle System
Classificação de assuntos	LCSH, AGROVOC, CDU

Quadro I. Bases de Dados para a conversão dos registros. Fonte: Elaborado pelos autores.

A partir dessas discussões acerca de quais bases de dados podem ser utilizadas para a conversão dos registros em Linked Data, o Quadro II apresenta um exemplo das possibilidades de estruturação de um registro de uma tese do Repositório da Institucional UNESP em Linked Data.

Elemento	**Valor**	**Base relacionada**	**Codificação**
dc.contributor.advisor	Santos, Plácida Leopoldina Ventura Amorim da Costa	-	-
	http://viaf.org/viaf/53615032364700997077 6/	VIAF	URI
	http://lattes.cnpq.br/7408791408049766	LATTES	URI
dc.contributor.author	Simionato, Ana Carolina	-	-
	orcid.org/0000-0002-0140-9110	ORCID	URI

	http://lattes.cnpq.br/9896600626524397	LATTES	URI
dc.date.issued	2015-03-10	-	W3-CDTF
dc.identifier.uri	http://hdl.handle.net/11449/123318	Repositório Institucional UNESP	URI
dc.language.iso	por	-	ISO 639-3 Codes
dc.publisher	Universidade Estadual Paulista (UNESP)	-	-
	http://isni.org/isni/0000000120969781	ISNI	URI
	http://viaf.org/viaf/122646643/	VIAF	URI
dc.source	Universidade Estadual Paulista (UNESP)	-	-
	http://isni.org/isni/0000000120969781	ISNI	URI
	http://viaf.org/viaf/122646643/	VIAF	URI
dc.subject	Catalogação		
	http://id.loc.gov/authorities/subjects/sh85020816	LCSH	URI
dc.subject	http://www.europeana.eu/portal/pt/record/08533/artifact_aspx_id_1063.html?q=Katsushika+Hokusai	Europeana	URI
dc.title	Modelagem conceitual DILAM: princípios descritivos de arquivos, bibliotecas e museus para o recurso imagético digital	-	-
dc.type	Tese de doutorado	-	-
dc.contributor.institution	Universidade Estadual Paulista (UNESP)	-	-
	http://isni.org/isni/0000000120969781	ISNI	URI
	http://viaf.org/viaf/122646643/	VIAF	URI
dc.relation.isPartOf	Modelo conceitual DILAM: integração entre arquivos, bibliotecas e museus	-	-
	http://www.ies.ufpb.br/ojs/index.php/ies/article/view/30902		URI

dc.rights.accessRights	Acesso aberto	-	-
	https://creativecommons.org/licenses/by/4.0/	creative commons	URI

Quadro II. Registro em Linked Data. Fonte: Elaborado pelos autores.

As possibilidades de ligação de dados conforme o exemplo do Quadro II revelou relacionamentos principalmente no controle de autoridades de pessoas, entidades coletivas e assunto, com as bases de dados VIAF, ORCID, LATTES, ISNI, LCSH. Em outros casos, o relacionamento de assunto com uma imagem que está na Europeana também é possível (especificidade do recurso em questão). Porém, outras bases podem ser usadas, como o Geonames para a identificação do local, bases para identificar um periódico, como Sherpa/Romeo, Diadorim ou Lucinéia ou, ainda, os relacionamentos poderiam ser utilizados DOI, administrado pela Crossref, entre outras possibilidades.

A partir da identificação da estrutura e das bases de dados que podem ser empregadas para a realização da conversão dos registros, é necessário pensar no processo completo para que um registro de um repositório digital seja disponibilizado nos princípios do Linked Data. Na próxima seção é exposta a proposta de um modelo para a realização deste processo.

5.2 Modelo para a disponibilização de dados de Repositórios Digitais em Linked Data

A estruturação dos recursos digitais em Linked Data deve seguir determinados esquemas e propostas, que visem inserir os benefícios do Linked Data nesses conjuntos de dados. No entanto, cada domínio possui particularidades e elementos que devem ser considerados para que esse processo seja eficiente e traga benefícios aos usuários.

No âmbito dos repositórios digitais, os sistemas utilizados e o modo como a conversão deve acontecer é essencial, para que ao final essa nova base de dados, com base nos princípios do Linked Data, possa contribuir com esse projeto. Torna-se, portanto, uma fonte importante de informações.

Deste modo, a Figura 1 mostra um esquema em que a partir de um conjunto de dados estruturados no DSpace possa se disponibilizar um conjunto de dados estruturados em RDF, seguindo os princípios do Linked Data em um banco de dados Virtuoso.

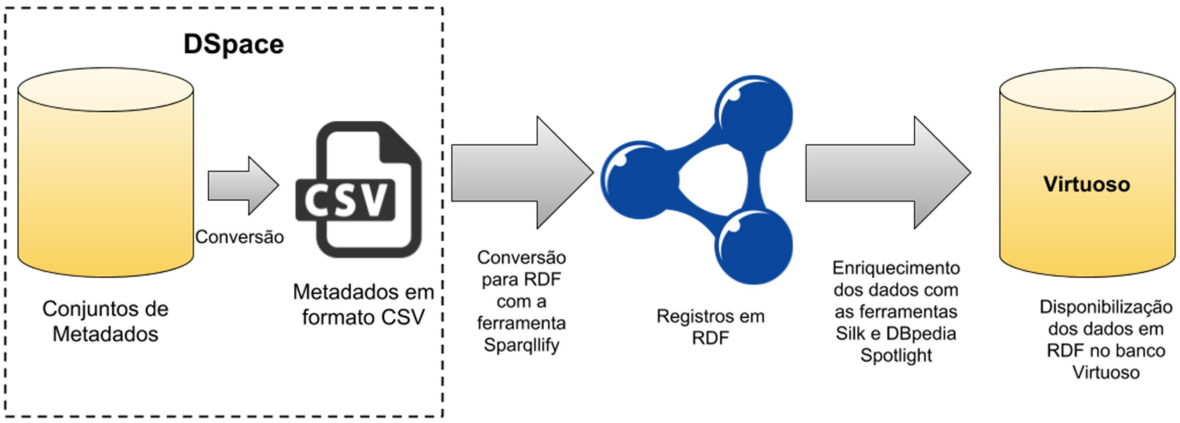

Figura 1. Modelo para disponibilização de dados nos princípios do Linked Data. Fonte: Elaborado pelos autores.

A Figura 1 aponta primeiramente os metadados no DSpace que estão originalmente inseridos em um banco de dados tradicional. Para que esses dados possam ser tratados e futuramente migrados para os princípios do Linked Data, primeiramente realiza-se a conversão dos metadados em formato Comma-Separated Values (CSV). O CSV é um formato em que os dados são estruturados e separados por meio de vírgulas. É bastante utilizado quando se tenta realizar processos automatizados de tratamento dos dados e conversão para outros padrões.

Em seguida, os metadados no formato CSV passam por um processo de conversão para RDF, por meio da ferramenta Sparqllify, que utiliza a linguagem SPARQL para criar os registros em RDF a partir do CSV. Essa ferramenta deve ser configurada de modo a explicitar como cada coluna do CSV será convertida para os grafos RDF.

Além disso, tal ferramenta possibilita que ontologias e outras bases de dados sejam relacionadas aos conjuntos de dados. As ontologias fornecem a estrutura que os dados devem seguir, e as outras bases de dados serão relacionadas aos registros, conforme a definição realizada no Quadro II. Vale destacar que o RDF é a base dos dados no Linked Data, cujo passo fundamental é converter os dados seguindo uma semântica formal adequada, promovida pelas ontologias e pelas demais bases de dados.

Posteriormente, deve ser realizado o processo de enriquecimento, em que serão utilizadas duas ferramentas. A primeira, o Silk, efetivará o enriquecimento ao promover um dos princípios do Linked Data, ao realizar a integração com outros datasets. Enquanto a segunda, o DBpedia Spotlight relacionará alguns conceitos, principalmente assuntos, com os seus registros no DBpedia. Este último processo auxilia na padronização de termos, além de oferecer mais possibilidades para os usuários navegarem pelos ambientes.

Ao final deste enriquecimento, os dados deverão ser disponibilizados em um banco de dados Virtuoso. Cria-se, então, uma base de dados que pode ser acessada por meio do SPARQL dos registros do repositório digital. Vale destacar que esse processo deverá ocorrer regularmente, para que o dataset esteja sempre atualizado com os novos registros inseridos no repositório digital.

Desta maneira, com esse esquema apresentado é possível ter ao final uma base de dados seguindo os princípios do Linked Data, dos dados oriundos de repositórios digitais. Ao se pensar na quantidade de informações que os repositórios digitais possuem no geral, verifica-se que ao obedecer a proposta desta pesquisa pode haver grande e importante aumento dos dados de Linked Data. Logo, favorece os usuários a encontrar produções intelectuais diversas.

6 CONSIDERAÇÕES FINAIS

O processo de conversão de dados para os princípios do Linked Data exige que uma série de etapas seja cumprida, especialmente para realizar a conversão para o RDF, de modo com que os dados estejam enriquecidos semanticamente. Neste sentido, há diversas bases que podem ser utilizadas para dar suporte a esse enriquecimento, ao mesmo tempo em que há diretrizes para a publicação de dados que podem suportar o processo de disponibilização dos datasets.

No âmbito deste trabalho foi abordado como os dados de repositórios digitais podem ser convertidos em conjuntos de dados seguindo os princípios do Linked Data. Os repositórios digitais contêm uma quantidade de dados estruturados extensa, principalmente de metadados, que se torna uma fonte importante e rica para a criação e a disponibilização de datasets.

Desta forma, o presente estudo demonstrou as possibilidades acerca deste processo de conversão. Identificaram-se bases de dados, vocabulários e padrões que devem ser adotados para que os dados gerados sejam enriquecidos semanticamente. O exemplo demonstrado é capaz de indicar a viabilidade da realização dos relacionamentos propostos, concomitantemente aponta como o padrão de metadados será representado e enriquecido para os princípios do Linked Data.

O trabalho apresentou, por fim, um modelo que reflete os passos que devem ser adotados no processo de disponibilização dos metadados de um repositório digital em Linked Data. Este esquema mostra os passos e as principais ferramentas que devem ser utilizadas para disponibilizar os dados. Inclusive, há uma fase que tange ao enriquecimento dos dados disponibilizados.

Portanto, esta pesquisa demonstra que a integração entre repositórios digitais e as tecnologias da Web Semântica permitem a disponibilização de dados em Linked Data. Fornece novos meios para a divulgação e à integração dos recursos na Web. Enquanto trabalhos futuros, busca-se implementar o trabalho realizado a propor meios de padronizar o processo de conversão de dados de repositórios digitais, em datasets de Linked Data.

7 REFERÊNCIAS

ALVES, R.C.V. and SANTOS, P.L.V.C.A. *Metadados no domínio bibliográfico*. Intertexto: Niterói, 2013.

ALVES, R.C.V. *Metadados como elementos do processo de catalogação*. [online]. Tese de doutorado, Universidade Estadual Paulista, Faculdade de Filosofia e Ciências de Marília, 2010. In: Repositório Institucional UNESP. Disponível em: <https://repositorio.unesp.br/handle/11449/103361> [Consulta: 6 de abril de 2019].

ALVES, R.C.V. *Web semântica*: *uma análise focada no uso de metadados*. [online]. Dissertação de mestrado, Universidade Estadual Paulista, Faculdade de Filosofia e Ciências de Marília, 2005. In: Repositório Institucional UNESP. Disponível em: <https://repositorio.unesp.br/handle/11449/93690> [Consulta: 6 de abril de 2019].

ARAKAKI, F.A. *Linked Data: ligação de dados bibliográficos*. [online]. Dissertação de mestrado, Universidade Estadual Paulista, Faculdade de Filosofia e Ciências de Marília, 2016. In: Repositório Institucional UNESP. Disponível em: <https://repositorio.unesp.br/handle/11449/147979> [Consulta: 6 de abril de 2019].

BACA, M. (ed.). *Introduction to Metadata: pathways to digital information*. Los Angeles, CA: Getty Information Institute, 1998.

BAKER, T. Libraries, languages of description, and Linked Data: A Dublin Core perspective. *Library Hi Tech*, 2012, vol. 30, nº 1, p. 116-133. ISSN 0737-8831

BAKER, T. et al. *Library Linked Data Incubator Group Final Report*. [online]. W3C Incubator Group Report, 2011. Disponível em: <http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/> [Consulta: 6 de abril de 2017].

BAPTISTA, A.A. and MACHADO, A.B. Um gato preto num quarto escuro: falando sobre metadados. *Revista de Biblioteconomia de Brasília*, 2011, vol. 25, nº 1, p. 77-90. ISSN 0100-7157

BERNERS-LEE, T. *Linked Data: Design Issues*. [online]. W3C, 2006. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html> [Consulta: 6 de abril de 2017].

BERNERS-LEE, T.; HENDLER, J. and LASSILA, O. The Semantic Web. *Scientific American*, may 2001, p. 29-37.

BYRNE, G. and GODDARD, L. The Strongest Link: Libraries and Linked Data. *D-lib Magazine*, 2010, vol. 16, nº 11/12. ISSN 1082-9873

CHOWDHURY, G.G. and CHOWDHURY, S. *Organizing information from the shelf to the web*. London: Facet Publishing, 2007.

CONARQ. *Diretrizes para a implementação de repositórios arquivísticos digitais confiáveis - RDC-Arq*. [online]. Rio de Janeiro, CONARQ, 2015. Disponível em: <http://www.conarq.arquivonacional.gov.br/images/publicacoes_textos/diretrizes_rdc_arq.pdf> [Consulta: 6 de abril de 2017].

CONFEDERATION OF OPEN ACCESS REPOSITORIES. *7 things you should know about…Linked Data*. [online]. 2017. Disponível em: <https://www.coar-repositories.org/activities/repository-observatory/second-edition-linked-open-data/7-things-you-should-know-about-open-data/> [Consulta: 6 de abril de 2017].

COYLE, K. and BAKER, T. *Guidelines for Dublin Core application profiles*. [online]. Dublin Core Metadata Initiative. 2009. Disponível em: <http://dublincore.org/documents/profile-guidelines/> [Consulta: 6 de abril de 2017].

DUFF, W. and MCKEMMISH, S. Metadata and ISO 9000 Compliance. *Information Management Journal*, 2000, vol. 34, nº 1. ISSN 0268-4012

DUVAL, E. et al. Metadata principles and practicalities. *D-Lib Magazine*, 2002, vol. 8, nº 4. ISSN 1082-9873

GONÇALEZ, P.R.V.A. *Repositórios arquivísticos digitais confiáveis: identificação de requisitos com ênfase no acesso à informação*. [online]. Tese de doutorado, Universidade Estadual Paulista, Faculdade de Filosofia e Ciências de Marília, 2017. In: Repositório Institucional UNESP. Disponível em: <https://repositorio.unesp.br/handle/11449/150028> [Consulta: 6 de abril de 2019].

LATIF, A.; BORST, T. and TOCHTERMANN, K. Testing the HathiTrust Copyright Search Protocol in Germany: A Pilot Project on Procedures and Resources. *D-lib Magazine*, 2014, vol. 20, nº 9/10, p.0-0. ISSN 1082-9873

LEITE, F.C.L. *Como gerenciar e ampliar a visibilidade da informação científica brasileira*: *repositórios institucionais de acesso aberto*. Brasília: IBICT, 2009.

LOD2. *LOD2*: *Creating Knowledge out of InterLinked Data*. [online]. Disponível em: <http://aksw.org/Projects/LOD2.html> [Consulta: 6 de abril de 2017].

LÓSCIO, B.F.; BURLE, C. and CALEGARI, N. *Data on the Web Best Practices*. [online]. W3C, 2017. Disponível em: <https://www.w3.org/TR/dwbp> [Consulta: 6 de abril de 2017].

MÉNDEZ RODRÍGUEZ, E. *Metadatos y recuperación de información*: estándares, problemas y aplicabilidad en bibliotecas digitales. Trea: Espanha, 2002.

MILLER, P. Interoperability: What Is It and Why Should I Want It? *Ariadne*, 2000, vol. 24. ISSN 1361-3200

POMERANTZ, J. *Metadata*. USA: The MIT press essential knowledge series, 2015.

RILEY, J. *Understanding metadata what is metadata, and what is it for?* EUA: NISO, 2017.

SEGUNDO, J.E.S. and CONEGLIAN, C.S. Web semântica e ontologias: um estudo sobre construção de axiomas e uso de inferências. *Informação & Informação*, 2016, vol. 21, nº 2, p. 217-244. ISSN 1981-8920

SAYÃO, L.F. Interoperabilidade das bibliotecas digitais: o papel dos sistemas de identificadores persistentes - URN, PURL, DOI, Handle System, CrossRef e OpenURL. *Transinformação*, 2007, vol. 19, nº 1, p. 65-82. ISSN 2318-0889

SIMIONATO, A.C. *Modelagem conceitual DILAM*: *princípios descritivos de arquivos, bibliotecas e museus para o recurso imagético digital*. [online]. Tese de doutorado, Universidade Estadual Paulista, Faculdade de Filosofia e

Ciências de Marília, 2015. In: Repositório Institucional UNESP. Disponível em: <https://repositorio.unesp.br/handle/11449/123318> [Consulta: 6 de abril de 2019].

SIMIONATO, A.C. *Representação, acesso, uso e reuso da imagem digital*. [online]. Dissertação de mestrado, Universidade Estadual Paulista, Faculdade de Filosofia e Ciências de Marília, 2012. In: Repositório Institucional UNESP. Disponível em: <https://repositorio.unesp.br/handle/11449/93646> [Consulta: 6 de abril de 2019].

VIDOTTI S. A.B.G. et al. Coleta automática para povoamento de repositórios digitais: conversão de registros utilizando XSLT. *Tendências da Pesquisa Brasileira em Ciência da Informação*, 2016, vol. 9, nº 2, p. 1-21. ISSN 1983-5116

WEIBEL, S. Metadata: the foundations of resource description. *D-Lib Magazine*, 1995, vol. 1, nº 1. ISSN 1082-9873

ZENG, M.L. and QIN, J. *Metadata*. New York: Neal-Schuman Publishers, 2008.

ZENG, M.L. and QIN, J. *Metadata*. 2. ed. New York: Neal-Schuman Publishers, 2016.