Proposta de um modelo de repositório digital utilizando conceitos da Web Semântica

Lucas Carmichael dos Santos de Oliveira

Faculdade de Tecnologia de Garça (FATEC) - lucas.oliveira212@fatec.sp.gov.br

Larissa Pavarini da Luz

Faculdade de Tecnologia de Garça (FATEC) - larissa.luz01@fatec.sp.gov.br

Resumo

A criação de repositórios digitais vem crescendo muito nos últimos anos, principalmente no contexto relacionado à disponibilização e publicação de dados na Web. Pensando neste aspecto, o presente trabalho tem por objetivo propor a criação de um repositório digital para a publicação de dados científicos com conceitos de Web Semântica. Este repositório terá como objetivo ampliar e disseminar as atividades de pesquisas desempenhadas por pesquisadores ou terceiros afins de gerar, divulgar e preservar a sua produção científica. Neste sentido, o trabalho apresenta uma revisão bibliográfica de natureza exploratória, partindo de conceitos iniciais, e uma proposta para construção de um Repositório Digital a partir da coleta de registros das diversas fontes que encontramos disponíveis na Internet, com o intuito de facilitar o uso dos dados publicados. Para isso, propõe-se a aplicação dos conceitos, fundamentos e boas práticas da Web Semântica e do Linked Data para permitir o enriquecimento, armazenamento e recuperação de informações de forma semântica e ligada, a partir de um vocabulário aberto.

Palavras-chave*: Web Semântica. Repositório Digital. Linked Data.*

Proposal of a digital repository model using semantic web concepts

Abstract

The creation of digital repositories has been growing a lot in the last years, mainly in the context related to the availability and publication of data on the Web. Thinking about this aspect, the present work aims to propose the creation of a digital repository for the publication of scientific data with concepts. Semantic Web This repository aims to expand and disseminate research activities performed by researchers or third parties in order to generate, disseminate and preserve their scientific production. In this sense, the paper presents an exploratory literature review, starting from initial concepts, and a proposal for the construction of a Digital Repository from the collection of records from the various sources we find available on the Internet, in order to facilitate the use of published data. For this, it is proposed to apply the concepts, fundamentals and good practices of the Semantic Web and Linked Data to allow the enrichment, storage and retrieval of information in a semantic and linked way, from an open vocabulary.

Keywords: *Semantic Web. Digital Repository. linked Data.*

1 Introdução

O Desenvolvimento da Web possibilitou o surgimento de um novo meio de interação e comunicação em sociedade que, além de absorver todas as mídias anteriores, permitiu o crescimento disponível e acessível em rede a todo o mundo. Como consequência, houve a necessidade de ferramentas capazes de encontrar entre inúmeros dados irrelevantes, uma informação precisa (PICKLER, M. E. V., 2007).

Considerada assim, a perspectiva de aumento na disponibilização de dados na Web estudantes, professores e pesquisadores da área científica tem o desafio de encontrar meios eficientes onde usuários possam acessar informações de seu interesse (SANTARÉM SEGUNDO et al. 2010).

A *World Wide Web Consortium* (W3C), recomendam soluções aos problemas de interoperabilidade entre sistemas de informação. Considera-se extensão da Web documentos a Web de dados, sendo que, atual Web parte do pressuposto que, a maior parte do conteúdo é destinada para interpretação dos seres humanos, não sendo facilmente entendida por sistemas computacionais, ou seja, máquinas. Entretanto a dificuldade abordada, faz-se a busca de estimular a organização dos dados na forma de relacionamentos conceituais, em redes, permitindo a atribuição de significados, para que sistemas computacionais e pessoas trabalhem juntas na recuperação semântica da informação (CRISTOVAO; FERNANDES, 2018).

Percebesse um grande volume de conteúdos na Web, encontrando-se em um cenário de acelerada expansão e sem muitos critérios para vinculação destas informações, podendo constatar dilemas nos quais as informações estão submetidas na Internet, ou seja, sem padrão para arquivamento dos dados ou implícita de métodos para lincagem dos dados e aplicabilidade semântica, possuindo difícil recuperação por parte dos agentes computacionais.

Segundo Tim Berners-Lee et al. (2001), pretende a evolução da Web de documentos para Web de dados por meio da semântica na qual informações disponibilizadas apresentem significado definidos, possibilitando o entendimento de softwares e de agentes computacionais, capazes de processar e interpretar automaticamente estes dados.

Complementando a visão de Tim Berners-Lee, Santarem Segundo (2010) define a Web de dados como um conjunto de padrões, que aplicados em uma massa de dados publicada na Web possa ser recuperada de forma semântica, ligando informações com o mesmo sentido e/ou significado, sem importar com sua estrutura sintática.

Vivemos um momento que precisamos da mudança, e que surjam com mais frequências inovações advindas da tecnologia da informação e comunicação, aplicando um papel preponderante neste cenário evolutivo. Modificar o modo que as tecnologias se relacionam é de importância fundamental (EVANS; WURSTER, 1999).

Uma tecnologia ferramental para se criar e tratar os assuntos focando em um problema que é a publicação de conteúdos científicos por parte de estudante, professores e pesquisadores em instituições acadêmicas, é os repositórios digitais, definido como sistemas de informações que armazenam, organizam e controlam arquivos digitais, permitindo a aplicação semântica e ligações destes dados, incorporando a facilidade da comunicação de agentes computacionais e seres humanos, colaborando com a proposta de Web Semântica (SAYÃO *et al*., 2009).

Os repositórios digitais (RDs), segundo o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT, 2018),

> [...] são bases de dados online que reúnem de maneira organizada a produção científica de uma instituição ou área temática. Os RD's armazenam arquivos de diversos formatos e ainda, resultam em uma série de benefícios tanto para os pesquisadores quanto às instituições ou sociedades científicas, proporcionando uma maior visibilidade aos resultados de pesquisas e possibilitando a preservação da memória científica de sua instituição.

A questão mais relevante para esta diversidade é a grande variedade de contextos, comunidades, objetivos e práticas ligadas à criação e funcionamento destes repositórios digitais. Estes podem ser concebidos como sistemas mundiais, cobrindo todos os assuntos e permitindo a qualquer pessoa colocar ou editar informações, ou como sistemas institucionais, ou sistemas por assuntos, destinados unicamente para utilizadores autorizados, com procedimentos de aprovação e de controle de qualidade (MARTINS; RODRIGUES; NUNES, 2008).

Assim, torna-se necessário clarificar quais os aspectos e características dos repositórios digitais que os diferenciam de base de dados, de sistemas de gestão de conteúdos, ou de outros que armazenam conteúdos digitais (HEERY; ANDERSON, 2005, p. 1-2): 1) os conteúdos são depositados em um repositório tanto pelo autor, pelo proprietário ou por um terceiro; 2) a arquitetura do repositório gera tanto conteúdo como metadados; 3) o repositório oferece um conjunto de serviços básicos mínimos, dentre os quais armazenar, organizar, recuperar e controle de acesso; 4) o repositório pode ser altamente sustentável e confiável, se bem gerido.

A cada diferente tipo de necessidade que uma comunidade define como prioridade existe diferentes tipos de soluções, uma delas é a de repositórios ou base de dados para atender problemas como grandes volumes de informações. Dito isso, a iniciativa de se criar diretrizes para publicações de dados vem de um movimento assistido monitorado nos últimos anos (NHACUONGUE *et al*., 2018).

Berners-Lee (2006) então define um conjunto de regras para a publicação de dados na Web, chamados de *Linked Data*. A intenção é que dados publicados seguindo um conjunto de regras sejam unificados em único espaço global de informações. Sendo elas: 1) Usar URI como nome para identificar as coisas; 2) Usar URIs HTTP para que as pessoas possam procurar por estes nomes e localizar o que deseja; 3) Fornecer informações úteis na recuperação dessas URIs, usando os padrões RDF e SPARQL que são a base da ligação de dados; 4) Incluir links para outros URIs, sendo possível descobrir mais informações sobre o assunto buscado. Regras estas que definem o princípio básico para publicação e conexão de informações através da Web.

Como citado anteriormente na quarta regra a ideia essencial de um modelo usando Linked Data através dos padrões RDF é demonstrar a ligação entre os dados através de triplas (um conjunto de Sujeito-Predicado-Objeto) utilizando o conceito de grafos e assim permitindo que determinado dado possa ser ligado a outro que fisicamente pode estar armazenado em outro arquivo, em qualquer outro lugar da web.

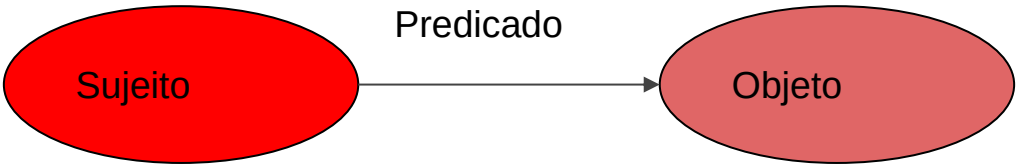

Figura 1 - Modelo de uma tripla RDF.
Fonte: autoria própria dos autores (2019).

O cenário do *Linked Data* ligado ao contexto acadêmico evolui a partir da disponibilização de novos conjuntos de dados que são relevantes para este contexto. A criação e publicação de conjuntos de dados é provocada pela necessidade de tornar acessíveis estas informações para consumo das pessoas. Processo este que é favorecer principalmente iniciativas e projetos com objetivo de reunir esforços e facilitar a troca de experiências, conhecimentos, ferramentas e vocabulários voltados para o Linked Data (NHACUONGUE *et al.*, 2017).

No contexto acadêmico existem iniciativas que seguem e disseminam os princípios do *Linked Data*. Iniciativas estas que tangem o propósito deste artigo se baseiam na *Linked Universities*, a qual é uma aliança de iniciativas que busca tornar seus dados públicos, por meio do *Linked Data*. Entretanto faz-se necessária a criação de vocabulários para que a integração de recursos na Web por meio do *Linked Data* seja possível.

2 Justificativa

De acordo com o contexto exposto, evidencia-se a necessidade do desenvolvimento de um futuro Repositório Digital de dados para controle de publicações científicas acadêmicas, possibilitando a disseminação de conteúdos relevantes de forma livre pela Web garantindo princípios da aplicabilidade semântica e do *Linked Data*. Observa-se que estas publicações pode ser realizado de maneira fácil e eficaz através de um Repositório Digital, trazendo benefícios com meios mais fáceis de publicações científicas, contribuindo com a expansão de conhecimento da comunidade acadêmica, proporcionando agilidade na recuperação da informação, reduzindo de forma significativa a perda das informações com passar do tempo.

3 Objetivo

O objetivo geral deste projeto de pesquisa é propor os passos iniciais para a criação de um repositório digital para as instituições acadêmicas. O intuito está em verificar a utilização da Web Semântica e do *Linked Data*, em conjunto, para aprimorar a forma como os usuários interagem com o sistema, permitindo a recuperação da informação.

A fim de que os objetivos sejam alcançados, buscam-se os seguintes passos: 1) Levantamento do problema; 2) Identificar a massa documental; 3) Mapear a estrutura para o repositório digital; 4) Estudar e utilizar tecnologias já existentes para construção do repositório (Ex.: DSpace CRIS); 5) Permitir a recuperação da informação, de forma semântica e ligada, em uso aos dados abertos.

4 Desenvolvimento

A pesquisa científica, até o presente momento, foi realizada por meio de revisão bibliográfica de natureza exploratória (ainda em andamento), além de ter sido definida a importância da análise dos dados já existentes.

Inicialmente, ao se analisar como são preservadas as publicações científicas, descobriu-se a existência de diversos repositórios (Ex.: SciElo), entretanto seus registros encontram-se privados, apenas para consulta de resumos. Além disso, percebemos que para se publicar e ter notoriedade relevante sobre o conteúdo elaborado é necessário arcar com custos monetários. Ao consultar outras bases como a do Google Schollar percebemos os fundamentos de ligação e aplicabilidade semântica dos dados por meio de vínculo de artigos. Estas informações levantadas esclarecem a importância de se desenvolver um repositório digital de modo que os dados sejam abertos à comunidade e livres de custos monetários para fins de incentivar a pesquisa acadêmica.

Os trabalhos de Dalziel (2005), e de Margaryan, Currier, Littlejohn e Nicol (2006), afirmam que repositórios digitais devem trazer bons resultados e boa acessibilidade, pois promover o compartilhamento e a reutilização de recursos de aprendizagem devem estar relacionados muito mais na comunidade do que no repositório digital em si, ou seja, nas atividades de aprendizagem mais do que nos conteúdos. Seu compartilhamento deve estar relacionado com as necessidades das pesquisas, e não (principalmente) pelo poder da tecnologia.

Instituições acadêmicas são lugares propícios para se desenvolver repositórios digitais úteis e bem sucedidos. Gerar serviço, onde os recursos da informação são ligados e organizados, com o propósito de serem de livre acesso à comunidade.

Desta forma, foram definidos métodos a serem seguidos para o desenvolvimento do repositório digital de dados: 1) Coleta dos registros: a coleta será por meio de um software desenvolvido para mineração, extração e análise de dados (estruturados, semiestruturados e não estruturados); 2) Normalização dos dados, ou seja, a definição do processo formal em que examina os conteúdos obtidos, com o objetivo de evitar anomalias antes de persistir registros no nosso repositório proposto; 3) Aplicação semântica: todo os estudo das diretrizes que serão necessárias para o mapeamento dos dados; 4) O estudo das tecnologias que serão utilizadas no repositório digital; 5) Fundamentação do Linked Data: uso de técnicas e boas práticas para a publicação de dados abertos e ligados; 6) Persistência de dados: armazenar, organizar e controlar os dados obtidos por meio de novas publicações e/ou da coleta de registros descrito no passo 1); e 7) Recuperação da informação: garantir a semântica e a ligação dos dados permitindo seu uso contínuo e eficaz.

Após a ideia apresentada, foi realizado um mapeamento com a ilustração dos pontos abordados acima. Para mais detalhes segue abaixo a figura:

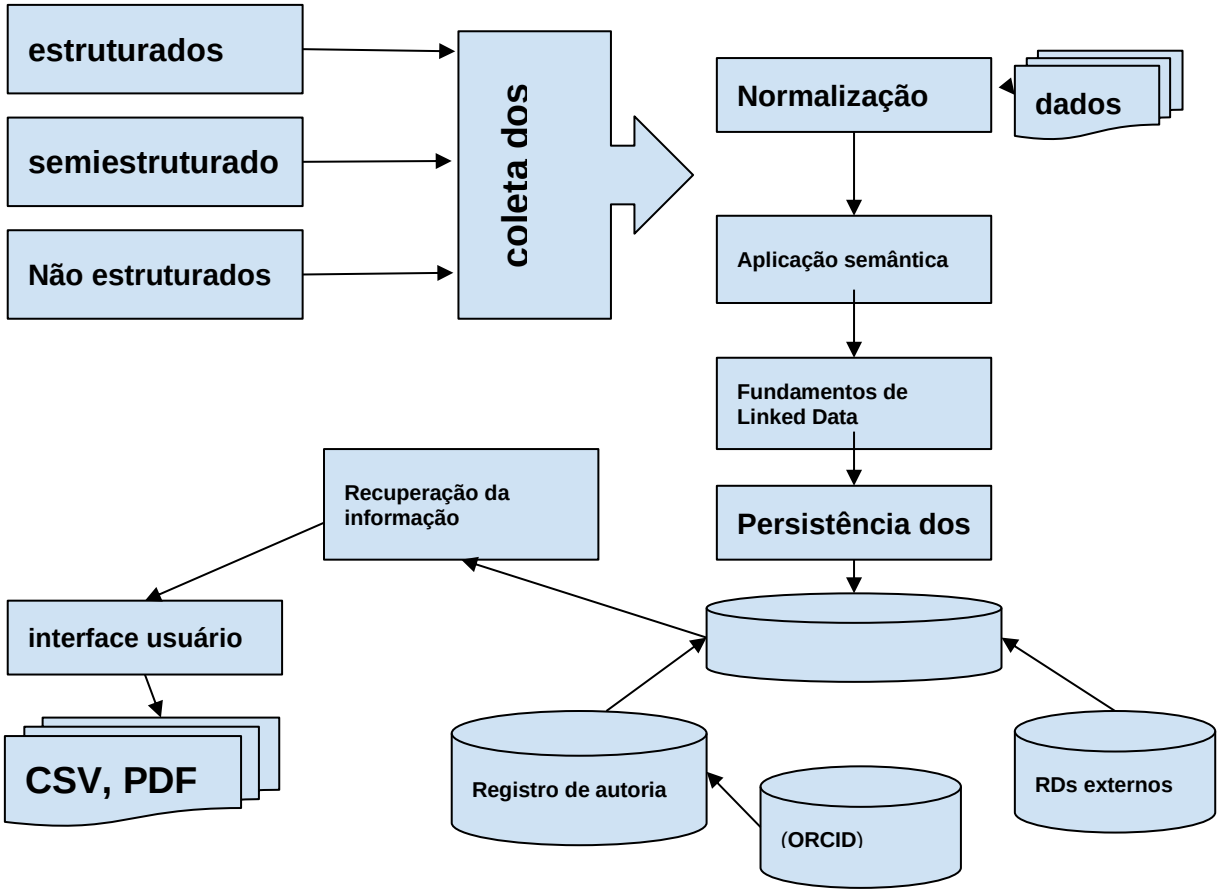

Figura 2 - Repositório Digital.
Fonte: autoria própria dos autores (2019).

Identificamos inicialmente a coleta das informações como sendo fator preponderante para a estruturação da criação da base de dados. Foi considerada então, três possibilidades de encontrar tais informações, que podem estar disponibilizadas como: a) **dados estruturados,** que existem em uma base sendo relacional, ou não que possamos ter o acesso facilitado; b) **dados semiestruturados,** que são todos os dados que não foram persistidos em uma base de dados, porém se encontram em estrutura de tags HTML possuindo certa semântica; e c) **dados não estruturados,** que são todos os dados contidos em documentos ou arquivos de diferentes formas, exemplo (PDF, DOC, ODT ou TXT), onde sua estrutura não é definida de forma similar entre arquivos, formando o não estruturamento de dados.

Diante disto, foi proposto o uso futuro do *OpenRefine* para coleta e transformações dos registros encontrados. A definição da ferramenta será descrita no próximo tópico 4.1 Ferramentas e Tecnologias.

A normalização da informação extraída no processo de coleta surgiu com o intuito de organizar e padronizar a massa de dados encontrada, aplicando semântica e os conceitos de *Linked Data* nas próximas etapas do processo, para permitir que agentes computacionais as identificassem no sentido das buscas e armazenamento das informações publicadas. Foi definida a persistência destas informações como de suma importância a povoar o repositório digital e, com as tecnologias que definiremos a seguir, garantir uma recuperação da informação de forma semântica e ligada às demais bases de dados abertas existentes.

Com bases em um levantamento documental preliminar realizado, alguns pontos importantes devem ser considerados no projeto inicial que embasa esta pesquisa, os quais são apresentados a seguir.

4.1 Ferramentas e Tecnologias

A etapa seguinte do projeto foi a definição de uma plataforma de software para a implantação do repositório digital pretendido.

Assim, após a análise entre alternativas de softwares de código aberto, bem como considerada a opção de desenvolvimento interno, foi escolhida a plataforma de software de código aberto DSpace-CRIS, a qual amplia as possibilidades de construção de repositório digital com características de um sistema CRIS. Tal escolha foi motivada pela necessidade de se ter uma plataforma que fosse capaz de gerenciar e de relacionar os dados utilizados e produzidos em uma pesquisa de um projeto específico, e assim abarcasse as produções científicas resultantes deste processo.

Como uma das funcionalidades importantes o DSpace-CRIS possibilitaria, na página do autor, vincular informações com o *Open Researcher* e o *Contributor ID*, as informações sobre o pesquisador e seus dados atrelados ao ORCID, além de reunir as produções desenvolvidas por ele. O vínculo com o ORCID permite que as informações cadastradas pelo indivíduo sejam inseridas no sistema, possibilitando localizar e escolher autores dos documentos inseridos pelo seu cadastro no ORCID (VIDOTTI *et al*, 2017), gerando uma das formas de enriquecer semanticamente o repositório.

Para cada um dos processos envolvidos na construção do repositório existem diversas ferramentas específicas. A partir do estudo em iniciação, identificamos algumas delas para que a pesquisa tenha desenvoltura. No primeiro contexto que envolve a identificação e descrição dos dados e o armazenamento e correção dos mesmo, envolve o uso do DSpace, como já citado anteriormente; o *OpenRefine* ferramenta que pode contribuir no processo de extração e a limpeza dos dados, uma vez que ela permite garantir a extração e transformação dos dados procedentes de diversos formatos, tal como descrito nos métodos.

Na etapa relativa à modelagem e ligação dos dados, a verificação do melhor vocabulário para descrever os dados com recursos RDF (*Resource Description Framework*), deve considerar os que já estão prontos na LOV (*Linked Open Vocabularies*, 2019), e verificar se existe a necessidade de gerenciá-los junto com ontologias, ou até mesmo criá-las para o projeto. No caso de dados de pesquisa científica, alguns vocabulários já são suficientes para classificar as definições de metadados. Para a publicação dos dados do repositório, ainda devem ser estudadas as possibilidades.

5 Resultados Esperados

Com a implantação do repositório Digital nas instituições acadêmicas, e uma vez povoado os dados verificam-se todas as questões de como serão tratados os processos semânticos, uma vez que a publicação dos dados vinculados está totalmente ligada a este contexto e à recuperação desses dados.

A criação do repositório proporciona diversas possibilidades de resultados, que podem ser desde a melhora da qualidade dos dados até, principalmente, a encontrabilidade da produção científica desenvolvida no meio acadêmico.

6 Conclusão

A implementação do repositório digital proposto nesse projeto, se encontra em sua fase inicial, e se pretende obter os resultados propostos anteriormente. O entendimento do cenário neste início de projeto é de vasto enriquecimento. Este repositório visa incentivar pesquisadores, professores e alunos a buscarem conteúdos para seus estudos, realizar publicações na própria plataforma, bem como correlacionar os dados persistidos no repositório digital com outras bases de dados abertas.

Serão apresentados em novos artigos futuros todos os processos de desenvolvimento do repositório digital proposto neste trabalho.

Como trabalhos futuros, recomenda-se o estudo mais aprofundado nas ferramentas e tecnologias abordadas neste trabalho, bem como sua forma de aplicação em repositórios digitais já implantados e em operação.

7 Referências

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. **The semantic web**. *Scientific American*, v.284, n.5, p.34-43, 2001.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked Data - **The Story So Far**. International journal on semantic web and information systems, v. 5, n. 3, p. 1–22, 2009.

CRISTOVÃO, H. M.; FERNANDES, J. H. C. Recuperação de informação em dados ligados: um modelo baseado em mapas conceituais e análise de redes complexas. **Transinformação**, v. 30, n. 2, p. 193-207, 2018. Disponível: http://dx.doi.org/ 10.1590/2318-08892018000200005. Acesso em: 15 set. 2019.

EVANS, P.; WURSTER, T. Getting Real About Virtual Commerce. **Harvard Business Review**, November-December 1999.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA (IBICT). **Repositórios Digitais**. Disponivel: http://www.ibict.br/informacao- para-a-pesquisa/repositorios-digitais#repositorios-brasileiro. Acesso em: 10 set 2019.

HEERY, R.; ANDERSON, S. **Digital Repositories Review**. Joint Information Committee. University of Bath. http://www.ukoln.ac.uk/repositories/publications/review-200502/digital-repositories-review-2005.pdf. Acesso em: 22 Set. 2019.

LINKED DATA. **Linked Data - Connect Distributed Data across the Web**. Disponível em: http://linkeddata.org/. Acesso em: 11 set 2019.

MARGARYAN, A.; CURRIER, S.; LITTLEJOHN, A.; NICOL, D. **Learning communities and repositories.** Disponível: https://www.gcu.ac.uk/cd-lor//learningcommunitiesreport.pdf. Acesso em: 22 set 2019.

MARTINS, A. B.; NUNES, M. B.; RODRIGUES, E. **Repositórios de informação e ambientes de aprendizagem: criação de espaços virtuais para a promoção da literacia e da responsabilidade social**. RBE - Rede de Bibliotecas Escolares Newsletter, Lisboa, n. 3, 2008. Disponível em: http://www.scielo.br/pdf/pci/v16n3/12.pdf. Acessado em: 22 Set. 2019.

NHACUONGUE, J. A.; ROZSA, V.; LIMA DUTRA, M. Linked Data e Ciência da Informação: diretrizes para a publicação de datasets institucionais abertos. **Biblios [online]**. 2018, n.73, pp.20-34. ISSN 1562-4730. Disponível em: http://dx.doi.org/10.5195/biblios.2018.429.

OPENREFINE. **A free, open source, powerful tool for working with messy data**. Disponivel em http://openrefine.org/. Acesso em: 22 Set. 2019.

PICKLER, Maria Elisa Valentim. **Web Semântica: ontologias como ferramentas de representação do conhecimento.** *Perspect. ciênc. inf.* [online]. 2007, vol.12, n.1, pp.65-83. ISSN 1981-5344. http://dx.doi.org/10.1590/S1413-99362007000100006. Acessado em 22 Set. 2019.

RODRIGUES, R., S.; TAGA, V.; VIERIA, E. M. F.; **Repositórios Educacionais: estudo preliminares para a Universidade Aberta do Brasil**. Disponível em: http://www.brapci.inf.br/_repositorio/2015/01/pdf_d7c05dcbb5_0027522.pdf. Acesso em 22 Set. 2019.

ROZSA, V.; DUTRA, M.; NHACUONGUE, J. Linked Open Data no contexto acadêmico: identificação e análise de vocabulários utilizados na academia e na pesquisa científica. **Brazilian Journal of Information Science: research trends.** v. 11, n. 3, 9 out. 2017.

SAYÃO, L. F.; SALES, L. F. Algumas considerações sobre os repositórios digitais de dados de pesquisa. **Informação & Informação**, v. 21, n. 2, p. 90-115, 2016a. Disponível em http://www.uel.br/revistas/ uel/index.php/informacao/article/viewFile/27939/20122. Acesso em: 14 set. 2019.

SANTAREM SEGUNDO, J. E. **Representação Iterativa: um modelo para Repositório Digital**. 2010. "bibliografia 140-150 f." Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília. 2010.

SENSO, J. A.; ARROYO MACHADO, W. (2018). La publicación en Linked Data de registros bibliográficos: modelo e implementación. **Revista Española de Documentación Científica**, 41 (4): e217. Disponível em: https://doi.org/10.3989/redc.2018.4.1535.

VIDOTTI, S. A. B. G. et al. **Coleta automática para povoamento de repositórios digitais: conversão de registros utilizando XSLT**. In: Encontro Nacional de Pesquisa em Ciência da Informação, 17., 2016, Bahia, 2016. **Anais...** Bahia: ANCIB; UFBA, 2016, p. 1-21. Disponível em: http://hdl.handle.net/11449/144718.

VIDOTTI, S. A. B. G. et al. **Repositório de dados de pesquisa para grupo de pesquisa: um projeto piloto**. Disponível em: http://enancib.marilia.unesp.br/index.php/xviiienancib/ENANCIB/paper/viewFile/388/932. Acesso em: 22 Set. 2019