O uso do Google OpenRefine na Padronização de metadados do Repositório Institucional da Universidade Federal da Bahia

Uillis de Assis Santos (UFBA) - uillisassis@hotmail.com

Kleber Carvalho Ferreira (UFBA) - kcf@ufba.br

Diana Paula de Oliveira Assis (Ifbaiano) - bibliotecariadiana@gmail.com

Gustavo Pinho Gomes dos Santos (UFBA) - gustavos@ufba.br

Adriene Marchiori (UFBA) - adrienemarchiori@yahoo.com.br

Resumo:

O trabalho relata o uso do software livre Google OpenRefine na padronização dos metadados do Repositório da Universidade Federal da Bahia (RI/UFBA). O procedimento foi realizado especificamente no metadado "dc.type", que possui a função de identificar a tipologia de cada documento. A intervenção foi necessária por dois motivos, a) identificação de uma desconfiguração; b) a não padronização conforme os metadados utilizados pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). Os problemas encontrados nos metadados não permitiam a importação dos arquivos depositados no RI/UFBA para os seguintes sistemas de informação: Banco Digital de Teses e Dissertações (BDTD), Portal Brasileiro de Publicações Científicas em Acesso Aberto (OASISBr), Red de Repositorios de Acesso Abierto a La Ciencia (La Referencia) e a Networked Digital Library of Theses and Dissertations (NDLTD), assim, impactando negativamente na visibilidade da produção acadêmica da universidade. Com a padronização dos metadados através do Google OpenRefine, foi possível corrigir mais de 21 mil arquivos de 395 coleções de diferentes Comunidades, além, de reestabelecer o envio dos arquivos para os sistemas de informação já citados. O relato ainda apresenta importantes contribuições para as equipes que administram repositórios, pois, demonstra a utilização de um software livre para a padronização de metadados em lotes, de forma rápida e segura. Ressaltando que a padronização proporciona também uma recuperação de arquivos de forma mais consistente por parte dos usuários do sistema e promove a visibilidade da instituição, dos autores e dos trabalhos depositados, ao mesmo tempo em que adota as diretrizes do Open Access.

Palavras-chave: *Google OpenRefine - Padronização de metadados; Repositório Institucional - Universidade Fedral da Bahia; Produção acadêmica - Visibilidade*

Eixo temático: *Eixo 6: Gestão de bibliotecas*

INTRODUÇÃO

O relato tem como objetivo apresentar o uso do software *OpenRefine* na padronização dos metadados do Repositório Institucional da Universidade Federal da Bahia (RI/UFBA), o procedimento foi realizado especificamente no metadado “dc. type”, que possui a função de identificar a tipologia de cada documento.

O *OpenRefine* foi criado em 2010 pelo Google “tem como objetivo principal tratar e limpar massas de dados sem que usuários necessitem conhecer com profundidade, termos e conceitos deste tipo de procedimento”

O trabalho se justifica por apresentar novas possibilidades para a correção dos metadados e demonstrar os impactos que os registros incorretos têm na visibilidade da produção do RI/UFBA.

Para contextualização do trabalho, faz-se necessário apresentar um pouco sobre o RI/UFBA e a importância do uso e padronização dos metadados. O RI/UFBA foi instalado em 2007, a princípio para disponibilizar as publicações da editora da UFBA. A partir de 2010, através da Portaria nº 024/2010, passou a ter o status institucional, com o propósito de divulgar e estabelecer o acesso a produção acadêmica desenvolvida no âmbito da Universidade.

Alinhada com as políticas de informação da UFBA, o Repositório possibilita também a preservação das produções depositadas em sua base, seja ela científica, cultural ou artística, ao mesmo tempo que adota as diretrizes do *Open Access*.

Para possibilitar a organização e visibilidade dos conteúdos armazenados em sua base, o RI/UFBA possui um padrão de metadados.

Os metadados desempenham um importante papel na identificação, localização e recuperação da informação nos Repositórios. Comumente metadados são definidos como “dados sobre dados” (RCAAP, 2019), no entendimento de Caplan (2003, p.3), “metadados são utilizados para significar informação estruturada sobre um recurso de informação de qualquer tipo de mídia ou formato”, porém, para que as informações estejam estruturadas, armazenadas e possam ser recuperadas, faz-se necessário que os repositórios possuam um padrão de metadados instalado em seu sistema.

> [...] padrões que estabelecem regras para definição de atributos (metadados) de recursos informacionais, para a) obter coerência interna entre os elementos por meio de semântica e sintaxe; b) promover necessária facilidade para esses recursos serem recuperados pelos usuários c) permitir a interoperabilidade dos recursos de informação. (ROSETO, 2003, p.59 apud GOMES, 2016, p.95)

Existem diversos tipos de padrões de metadados que são utilizados de acordo com as suas funções, Gillilard (2008) define cinco tipos de metadados: administrativo, descritivo, preservação, técnico e uso. O software utilizado pelo RI/UFBA é o DSPACE, que já vem com o padrão de metadados Dublin Core instalado no sistema. O Dublin Core é caracterizado como padrão de metadado descritivo, assim, possui a função de identificar e descrever objetos, sendo digitais ou não, como por exemplo vídeos, sons, imagens, textos e sites na web.

Os padrões de metadados são fundamentais também para possibilitar a interoperabilidade entre sistemas, um exemplo são os sistemas de informações que proporcionam uma maior visibilidade para os repositórios: Banco Digital de Teses e Dissertações (BDTD), Portal Brasileiro de Publicações Científicas em Acesso Aberto (OASISBr), Red de Repositorios de Acesso Abierto a La Ciencia (La Referencia) e a Networked Digital Library of Theses and Dissertations (NDLTD). Estes sistemas recebem as produções acadêmicas dos Repositórios: teses, dissertações, livros, capítulos de livros, artigos de periódicos, artigos de eventos, trabalhos de conclusão de curso e relatórios de pesquisa. A BDTD e o OASISBr recebem a produção acadêmica a nível nacional, LA Referencia recebe a produção acadêmica da América Latina e a NDLTD recebe as produções a nível mundial.

RELATO DA EXPERIÊNCIA

Durante o processo de atualização do DSPACE da versão 3.2 para 5.7, a equipe identificou uma incompatibilidade no metadado “dc. type”. Além de não estar em conformidade com os metadados utilizados pelo IBICT, a inconformidade não permitia a importação de uma cópia das produções do RI/UFBA para os seguintes sistemas de informação: BDTD, OASISBr, LA Referencia e a NDLTD. A tabela a seguir apresenta a diferença entre os valores de “dc. type” do RI/UFBA e do IBICT:

IBICT “dc. type”	RI/UFBA “dc. type”
Tese	Trabalhos finais e parciais de curso: Teses de Doutoramento (defendida e aprovada por banca especializada)
Dissertações	Trabalhos finais e parciais de curso: Dissertações de Mestrado (defendida e aprovada por banca especializada)
Livros	Produção bibliográfica: Livros
Capítulo de Livro	Produção bibliográfica: Capítulo de Livro
Artigo de Periódico	Produção bibliográfica: Artigos completos publicados em periódicos
Artigo de Evento	Produção bibliográfica: Trabalhos publicados em anais de eventos
Trabalho de Conclusão de Curso	Trabalhos finais e parciais de curso: Trabalhos de Conclusão de Graduação
Relatório de Pesquisa	Produção técnica: Relatório de pesquisa

Fonte: elaboração do autor.

As ações para a atualização do RI/UFBA somente poderiam ser retomadas após as correções destes metadados.

A princípio foi realizada uma reunião no Campus de Ondina da UFBA na Superintendência de Tecnologia e Informação (STI), com a intenção de elaborar um planejamento para solucionar o problema. Estiveram presentes na reunião a equipe que administra o RI/UFBA e um funcionário da STI.

O uso do software *OpenRefine* foi indicado pelo funcionário da STI para ser aplicado na correção dos metadados, a versão utilizada foi a 2.8, que suporta arquivos como CSV (padrão de exportação do DSPACE), TSV, Excel, JSON, XML e RDF.

Inicialmente algumas dificuldades foram encontradas para utilização do software, pois, o manual cedido pelo IBICT não constava com todas as informações necessárias para as configurações, assim, resultando em mais duas reuniões para a efetivação da configuração.

As correções foram concretizadas no prazo de dois meses entre 04/04/18 à 04/06/18 e foram corrigidos mais de 21 mil arquivos de 395 coleções de diferentes Comunidades. Uma limitação do software é que somente permitia editar até 500 arquivos por vez.

Para demonstrar a relevância da ação executada, a equipe do RI/UFBA registrou o antes e o depois das quantidades de arquivos encontrados nos sistemas de informação. Apenas da NDLTD que não foi possível obter os dados.

Sistemas de informação	Arquivos antes	Arquivos depois
BDTD	1.633	9.211
OASISBr	4.747	20.815
LA REFERENCIA	11.825	16.939

Fonte: elaboração do autor.

CONSIDERAÇÕES FINAIS

Os objetivos foram alcançados com a padronização dos metadados do RI/UFBA e o reestabelecimento da migração das produções acadêmicas para as bases da BDTD, OASISBr, LA Referencia e a NDLTD.

O uso do *OpenRefine* não acarretou em nenhum custo para a universidade, pois, tratava-se de um software gratuito, corroborando assim com as recomendações do *Open Access*

O relato ainda apresenta importantes contribuições para as equipes que administram repositórios, pois, demonstra a utilização de um software para a padronização de metadados em lotes, de forma rápida e segura. Ressaltando que a padronização proporciona também, uma recuperação de arquivo de forma mais consistente por parte dos usuários do sistema, promove a visibilidade da instituição, dos autores e dos trabalhos depositados em conformidade com as diretrizes do *Open Access.*

Lembrando que as correções foram realizadas no metadado “dc. type”, que identifica o tipo de documento, porém, o *OpenRefine* de acordo com o seu manual, pode ser utilizado em outros metadados como título, autor, orientador, acesso, data de publicação, entre outros, além de permitir a atomização de dados, eliminação de duplicações e tratamento de valores múltiplos. As funcionalidades descritas podem ser utilizadas para contribuir de forma significativa no controle da qualidade dos dados depositados nos repositórios.

REFERÊNCIAS

BIBLIOTECA DIGITAL DE TESES E DISSERTAÇÕES (BDTD). Disponível em: http://bdtd.ibict.br/vufind/. Acesso em: 22 mar. 2018

CAPLAN, Priscila. **Metadata fundamentals for All librarians**. Chigaco: Americam Librarian Association, 2003. Disponível em: https://books.google.com.br/books?id=yt2863FismcC&pg=PA1&hl=ptBR&source=gbs_toc_r&cad=#v=onepage&q&f=false. Acesso em: 2 fev. 2019.

GILLILAND, Anne J. Setting the stage. In: **Introduction to metadata**. Califórnia: The Getty Research Institute, 2008. Disponível em: http://d2aohiyo3d3idm.cloudfront.net/publications/virtuallibrary/0892368969.pdf. Acesso em: 3 fev. 2019.

GOMES, Fábio Andrade. **Padronização de metadados na representação da informação em repositórios institucionais de universidades federais brasileiras**. 2015. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal da Bahia, Salvador, 2015. p. 277. Disponível em: https://repositorio.ufba.br/ri/handle/ri/18950. Acesso em: 3 fev. 2019.

GOOGLE OPEN REFINE. Disponível em: http://openrefine.org/download.html. Acesso em: 23 mar. 2018.

NETWORKED DIGITAL LIBRARY OF THESES AND DISSERTATIONS (NDLTD). Disponível em: http://www.ndltd.org/. Acesso em: 22 mar. 2018.

PORTAL BRASILEIRO DE PUBLICAÇÕES CIENTÍFICAS EM ACESSO ABERTO (OASISBr). Disponível em: http://oasisbr.ibict.br/vufind/. Acesso em: 22 mar. 2018

RED DE REPOSITORIOS DE ACESSO ABIERTO A LA CIENCIA (LA REFERENCIA). Disponível em: http://www.lareferencia.info/pt/. Acesso em: 22 mar. 2018.

RILEY, Jenn. **Understanding metadata: what is metadata, and what is it for?** Baltimore: NISO, 2017. Disponível em: https://groups.niso.org/apps/group public/download.php/17446/Understandin%20Metadata.pdf. Acesso em: 5 fev. 2019.

REPOSITÓRIOS CIENTÍFICOS DE ACESSO ABERTO DE PORTUGAL (RCAAP). Disponível em: http://projeto.rcaap.pt/index.php/lang-pt/component/quickfaq/5-processodedeposito-auto-arquivo/55-o-que-sao-metadados. Acesso em: 5 fev. 2019.