DEPURAÇÃO DOS METADADOS NO REPOSITÓRIO DIGITAL DA FUNDAÇÃO GETULIO VARGAS

Bacha, Marcia Nunes[1]
Almeida, Maria do Socorro G. de[2]
Carvalho, Kelly M. Ayala de[3]
[1]FGV/Sistema de Bibliotecas/marcia.bacha@fgv.br
[2]FGV/Sistema de Bibliotecas/maria.socorro@fgv.br
[3]FGV/Sistema de Bibliotecas/kelly.ayala@fgv.br

RESUMO

Neste trabalho apresentamos o projeto de depuração de metadados realizado no Repositório Digital da Fundação Getulio Vargas (FGV). Na implantação do Repositório, o maior foco era a inserção dos documentos para torná-lo robusto e chamar a atenção da Instituição para os objetos digitais produzidos. Como a entrada dos documentos era feita pelos próprios autores, ou por migrações de sistemas acadêmicos, os dados entravam sem critérios estabelecidos e sem metadados significativos ou com estes preenchidos de forma incorreta. O projeto em questão visou criar padrões para entrada de dados e, principalmente, instruções de trabalho que normalizam e melhoram a qualidade dos metadados ao inserir os documentos no Repositório. Na primeira fase foram realizados levantamentos das ocorrências de inconsistências na plataforma. Após este processo e diante dos resultados, foram desenvolvidas estratégias e ações para o acerto dos registros. A adoção de um vocabulário controlado como forma de normalizar os assuntos, e a criação do glossário para definir os tipos de materiais existentes na plataforma, assim como implantação de uma política para entrada dos documentos, fizeram parte da metodologia adotada no projeto. Estas medidas de normalização e padronização tornaram o Repositório mais estruturado e homogêneo e o resultado foi a melhoria da visibilidade e interoperabilidade dos dados.

Palavras-Chave: Repositório digital. Metadados. DSpace. *Dublin Core*.

INTRODUÇÃO

O Repositório Digital FGV [1]é composto por um portal de periódicos (Periódicos Científicos e Revistas FGV) e um repositório institucional (Comunidades FGV), sendo que o primeiro sistema desenvolvido com o Open Journal System (OJS) e o segundo com o DSpace. Assim, a FGV busca atender a publicação de conhecimento por meio de periódicos e a preservação e disseminação pelo repositório. O presente trabalho está dedicado ao "Comunidades FGV", que mesmo com mais de dez anos nunca havia passado por um processo de análise dos seus metadados, voltados especificamente para a qualidade.

Entende-se que os metadados são dados estruturados que suportam funções associadas do objeto designado, conforme designação de Greenberg (2005). O padrão de metadados

[1] http://sistema.bibliotecas-bdigital.fgv.br/

Dublin Core é comumente utilizado em bibliotecas digitais para organizar a informação e sua recuperação efetiva através das pesquisas por meio das buscas. A qualidade na descrição dos metadados e sua padronização são fundamentais para facilitar o acesso na plataforma e a interoperabilidade dos dados. Além disso, Steve O'Connor (2016) acredita que a gramática presente nos repositórios tem crescido e sofrido mudanças, sendo necessária a intervenção do bibliotecário a fim de reexaminar, renovar e revitalizá-la, pois não é possível deixar este processo de evolução em modo automático, ou seja, sem um controle das transformações que ocorrem.

Em 2006, ano em que foi implantado o Sistema, foram migrados documentos provenientes dos Sistemas Acadêmicos da FGV e de bases de dados geradas pelas próprias unidades.

Neste período, os documentos entravam no DSpace sem nenhuma padronização e não passavam pelo fluxo de postagem criado pela equipe da Biblioteca. Diversos campos ficaram sem preenchimento ou foram preenchidos incorretamente. Somente a partir de 2014 a padronização dos metadados no Repositório se tornou parte intrínseca na rotina da equipe do Repositório Digital FGV.

O objetivo deste trabalho é relatar o processo de análise e depuração dos metadados dos documentos presentes no acervo do "Comunidades FGV".

METODOLOGIA E ETAPAS DO PROJETO

A fase inicial do projeto consistiu-se no levantamento e identificação, em planilhas Excel, das inconsistências nos registros e seus metadados.

A Equipe de TI desenvolveu um script onde foram escolhidos metadados fundamentais e aplicadas as regras para validação nesses campos. Duas ações foram realizadas: 1) Listar o conteúdo preenchido nesses campos, dividindo-os em planilhas, para facilitar a comparação e acerto de dados; 2) Criticar a existência do metadado. Como exemplo podemos citar o metadado dc.subjetc LCSH, metadado obrigatório em nossos itens. Para este caso criou-se uma regra para filtrar os metadados não presentes, gerando listagens com informações que identificassem o registro (lista de handles) e quantidade de ocorrências em cada coleção.

Neste cenário, foi possível identificar as seguintes situações:

- Autores e assuntos escritos de diversas formas;
- Registros sem arquivos associados;
- Registros onde não há o vocabulário controlado preenchido;

- Divergência, falta de padronização e ausência de preenchimento dos metadados dc.type e dc.abstracts.

O trabalho contou com profissionais de TI e bibliotecários do Repositório Institucional da FGV e terceirizados. Atualmente, após a publicação dos documentos pelos responsáveis da coleção, os bibliotecários do Repositório verificam se faltam informações, como por exemplo resumos, fontes de citação, notas, e também padronizam os autores e inserem os termos do vocabulário controlado

RESULTADOS E DISCUSSÕES

Foram revistos ao todo 10.113 ocorrências de assuntos que precisaram de revisão, 11.735 registros para a revisão e/ou inclusão do tipo de material, 2.890 registros de teses e dissertações com metadados inexistentes e com possíveis erros como: orientador inexistente, ausência de tópicos de alunos (keywords), inexistência dos termos controlados do vocabulário da *LCSH*[2], falta de resumo, resumo composto apenas pelo termo *None* e nenhum arquivo associado ao registro.

Após o levantamento, dois itens chamaram atenção e mereceram uma medida de ação imediata, foram eles os metadados de assuntos e os de tipo de material. O tipo de material sofreu uma revisão, tendo-se como base o novo glossário do Repositório Digital da FGV, onde os documentos se enquadrariam, gerando uma padronização e uniformidade ao metadado dc.type e além disto, documentos onde este metadado era inexistente foram estudados e atribuídos a um tipo documental.

Para os assuntos, determinou-se que o vocabulário controlado adotado seria a do Bibliodata[3], ou seja, a pesquisa para preenchimento do metadado assunto foi realizada na lista autoridades deste, com o objetivo de verificar se os autores estão padronizados, e criar um cabeçalho de assunto para o documento. Com relação ao projeto, como se tratava de um passivo, os cabeçalhos já constavam no documento. Neste caso, procedeu-se à busca do cabeçalho adotado no Catálogo de Autoridades para verificar se estava de acordo com o estabelecido. Caso o assunto adotado não constasse na Lista de Autoridades da Rede Bibliodata, seria realizada uma pesquisa em fontes específicas, como a *Library of Congress Subject Headings* da (LC) e a Biblioteca Nacional (BN) e este termo seria inserido, passando a compor o Catálogo de Autoridades.

[2] *Library of Congress Subject Heading*

[3]Rede Bibliodata é uma rede de bibliotecas com a finalidade de promover a catalogação cooperativa, o compartilhamento de registros bibliográficos e a disseminação dos acervos das bibliotecas brasileiras

Os registros de teses e dissertações com metadados considerados essenciais (orientador, resumo e palavras-chave), porém inexistentes, foram verificados e preenchidos caso o trabalho fornecesse a informação suprimida na catalogação. Foi verificado que, por não possuírem folha de rosto ou outra fonte de informação, trabalhos antigos não continham informações de orientador e, em alguns casos, não havia ocorrência de resumos.

Os assuntos do vocabulário controlado, após serem listados, foram conferidos e durante o processo de padronização, foram encontrados nas autoridades do repositório problemas de grafia, uso de termos não autorizados, e ordem incorreta dos subcabeçalhos. Além disto, aproveitou-se para efetuar a mudança dos termos para o novo acordo ortográfico. Após este trabalho de varredura e acerto, iniciou-se o processo de manutenção do vocabulário controlado (dc.subject.bibliodata).

Para manter a qualidade e evitar possíveis erros no metadado de tipo de material, foi implantando na plataforma uma caixa de seleção (lookup), utilizando-se os termos do novo glossário, onde eles são escolhidos de forma padronizada.

CONSIDERAÇÕES FINAIS

Visando promover tal avanço e controle**,** além de haver em mente o aumento da precisão na busca do repositório institucional e interoperabilidade dos dados, deu-se início ao projeto de depuração da base de dados dos registros contidos no Repositório Digital FGV.

Após a aplicação deste projeto, os metadados que foram migrados para plataforma por importação, ou seja, sem padrão pré-definido, foram depurados e se enquadraram no esquema de padronização de dados adotados pelo Repositório Digital.

O controle e manutenção deste projeto são primordiais para que não haja necessidade de retrabalho e permitir uma continuidade da homogeneidade e consistência formada na padronização dos tipos documentais, dos assuntos e dos autores existentes no DSpace.

Em suma, o projeto trouxe como resultado:

- A possibilidade do acesso à informação com maior rapidez e melhor qualidade;
- O incremento da interoperabilidade dos dados com outros repositórios, como o Google Scholar, e outros indexadores científicos;
- Ampliação da visibilidade dos documentos nacional e internacionalmente;
- Aumento da eficiência no processo de exportação dos registros para o projeto da Rede de Pesquisa da FGV.

Atualmente, a ênfase do Sistema de Bibliotecas FGV é aumentar o número de documentos publicados no Repositório, promovendo ações para incrementar como:

divulgação através de newsletter, estatísticas mensais e continuar o processo de sensibilização da comunidade interna a fazer novas submissões.

REFERÊNCIAS

BACHA, M. N.; ALMEIDA, M. S. G. Construindo a Biblioteca Digital da FGV: estudo de caso. In: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 17., 2012, Gramado. **Anais eletrônicos**... Gramado: UFRGS, 2012. Disponível em: <http://www.snbu2012.com.br/anais/pdf/4QJQ.pdf>. Acesso em: 15/03/2017.

Fundação Getulio Vargas. Sistema de Bibliotecas. **Estatísticas do Repositório Digital FGV**. Disponível em: <http://www.fgv.br/biblioteca/relatorio/index.html>. Acesso em: 15 mar. 2017.

GREENBERG, J. Metadata and the World Wide Web. In: Encyclopedia of Library and Information Science. New York: Marcel Dekker, 2005. p. 1876-1888. Disponível em < http://www.ils.unc.edu/mrc/pdf/greenberg03metadata.pdf> Acesso em: 12 mar. 2013.

JAIN, P. New trends and future applications/directions of institutional repositories in academic institutions. **Library Review**, [Reino Unido], v. 60, n. 2, 2011, p. 125-141. Disponível em: <http://dx.doi.org/10.1108/00242531111113078>. Acesso em: 10 fev. 2017.

O'CONNOR, S. Stating the problem: the grammar of repositories. **Library Management**, [Austrália], v. 37, n. 4-5, 2016, p. 210-220. Disponível em: <http://dx.doi.org/10.1108/LM-01-2016-0007>. Acesso em: 10 fev. 2017.

Rede Bibliodata. **Quem somos**. Disponível em: <http://bibliodata.ibict.br/>. Acesso em: 15 mar. 2017.

SAYÃO, L. F. Afinal, o que é biblioteca digital?. **Rev. USP**, São Paulo, n. 80, fev. 2009 . Disponível em: <http://www.revistas.usp.br/revusp/article/view/13709>. Acesso em: 18 jun. 2012.