cadernos BAD

Biblioteca Digital Brasileira de Teses e Dissertações: ações para melhoria na qualidade dos dados

Diego José Macêdo
Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
diegomacedo@ibict.br

Milton Shintaku
Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
shintaku@ibict.br

Tainá Batista de Assis
Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
taina@ibict.br

Washington L. R. de Carvalho Segundo
Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
washingonsegundo@ibict.br

Ronnie Fagundes de Brito
Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
ronniebrito@ibict.br

Resumo

Sistemas que implementam os preceitos dos Arquivos Abertos têm na interoperabilidade o ponto forte por possibilitar o intercâmbio de informação. A Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), primeiro sistema de informação brasileiro a implementar os Arquivos Abertos no Brasil, coleta automaticamente metadados de 104 provedores de dados que nem sempre apresentam-se normalizados. Assim, por meio de uma pesquisa empírica e experimental, desenvolveu um método e ferramentas (como crosswalks), que apoiam a qualidade dos dados proveniente dos provedores de dados que compõe a base da BDTD. Isso incrementou a quantidade de registros que apresentam padronização no preenchimento dos campos e normalização nos conteúdos dos campos, ofertando à comunidade usuária da BDTD uma base de dados mais apropriada às ferramentas de busca e descoberta.

Palavras-chave: Qualidade de metadados, Mapeamento, Crosswalk, BDTD.

Introdução

Em 2012, a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD) completou 10 anos sendo uma das primeiras redes brasileiras a implementar os preceitos dos Arquivos Abertos. Isso possibilitou o acesso às teses e dissertações, em texto completo, via web, de grande parte dos programas brasileiros de pós-graduação. Por se tratar de uma rede baseada nos arquivos abertos, compõe-se de provedores de dados - as bibliotecas digitais locais de teses e dissertações-, um sistema de coleta de metadados (*harvesting*) e um portal de serviços consolidados que também é um agregador. Os dados são provenientes de sistemas independentes, o que resulta em diferentes padronizações locais, tornando mais suscetíveis à ocorrência de variações em seus conteúdos, mesmo nos campos que possuem obrigatoriedade de normalização. Essas variações influenciam a recuperação da informação, dada a indexação automática dos metadados e podem resultar em indicadores inconsistentes.

Objetivo(s)

Deste modo, o presente estudo tem por objetivo apresentar os resultados de avaliação dos metadados descritivos da BDTD passíveis de normalização, os quais foram submetidos a algoritmos desenvolvidos para melhoria da qualidade dos dados agregados. Com isso, visa-se apoiar estudos voltados à intersecção da Ciência da Informação e da Ciência da Computação, principalmente no conteúdo referente às questões de recuperação da informação.

Metodologia

Contrastando com abordagem das ciências puras, a computação se caracteriza pelos estudos mistos de pesquisa e desenvolvimento, principalmente pelo alinhamento da disciplina à tecnologia aplicada. De característica empírica e experimental, o presente estudo se aproxima do que Wazlawick (2008) classifica de apresentação de produto virtual, visto que o autor considera a Ciência da Computação como a Ciência do Artificial, em oposição ao mundo real das Ciências Naturais. Assim, a presente pesquisa se baseou na coleta de metadados provenientes de sistemas de gestão de teses e dissertações, via protocolo *Open Archives Initiative - Protocol Metadata Harvesting* (OAI-PMH), com a aplicação de filtros para determinar variações, erros de preenchimento e grau de normalização.

Resultado(s) e Discussão

A BDTD interopera com 104 provedores de dados, desenvolvidos com tecnologias diversas, onde há destaque para as bases construídas com o Sistema de Publicação Eletrônica de Teses e Dissertações (TEDE) e o DSpace, com 84 e 15 provedores respectivamente. Nota-se que as instalações com DSpace são, em sua totalidade, repositórios institucionais (RIs) que disponibilizam outros tipos de documentos científicos, principalmente artigos científicos. Quanto à interoperabilidade, o sistema de coleta atual opera com diversos esquemas de metadados, convertendo-os, via transformações crosswalk, ao padrão de metadados de operação da BDTD. Nesse ponto, tem-se 92 bibliotecas utilizando o MTD2-BR, seis utilizam o DSpace Intermediate Metadata (DIM), cinco o Resource Description Framework (RDF) e o MarcXML possui apenas uma participante. Destaca-se o fato da não utilização do Dublin Core (DC) não qualificado, principalmente pela complexidade descritiva das teses e dissertações, as quais requerem um detalhamento descritivo que o DC simples não consegue ofertar, impedindo a manipulação necessária para melhoria da qualidade dos dados.

Para o processo de indexação, os registros coletados são convertidos automaticamente para o padrão adotado pela BDTD. Tarefa que, tecnicamente, é efetuada via aplicação de Crosswalks, estes construídos sobre transformações XSLT (Extensible Stylesheet Language Transformations). Com isso, viabiliza-se maior flexibilidade à rede, pois é permitido que cada biblioteca local adote o padrão de metadados mais apropriado ao seu contexto tecnológico e informacional. Da mesma forma, ocorrem mapeamentos que normalizam e padronizam o conteúdo de alguns campos como, idioma, tipo do documento, grau e instituição de defesa. Entre tantas expressões equivalentes encontradas, a avaliação revelou variações no campo Grau, tal como: Mestre, mestrado, mestrado em <nome do programa>. O campo de instituição de defesa apresenta ainda maior variação, pois são identificados preenchimentos com nome por extenso, siglas ou abreviaturas diversas e houve, portanto, necessidade de execução de algoritmos de normalização em grande número de casos. Campos como tipo de documento e idioma também sofreram transformações de conteúdo para que se alinhassem às orientações das diretrizes DRIVER.

Após análise dos resultados obtidos, foi possível o desenvolvimento de ferramentas que ajustaram os dados coletados. Em uma via recursiva de refinamento, aplicaram-se ferramentas de ajuste; efetuou-se nova análise, e esse processo permitiu uma melhor acurácia dos dados coletados. Assim, completou todo ciclo que visa alcançar refinamento da qualidade dos dados na base consolidada.

Conclusão

Constata-se a necessidade de processamento para melhoria da qualidade de dados em redes heterogêneas, composta por sistemas que operam com formatos de metadados diferentes. O processo adotado na presente pesquisa encontra apoio no estudo de Stupmf e McDonnell (2004), que indica como possível solução para problemas de acurácia de metadados o uso de ferramentas automatizadas. Nota-se que a tecnologia apoia não apenas a infraestrutura, possibilitando uma maior flexibilização aos provedores de dados, mas também o tratamento da informação. Com isso, torna-se mais eficaz a melhoria da disseminação da

informação.

Referências

STUMPF, S.; McDONNELl, J. (2004) - Sharing Metadata – problems and potential solutions. Em: *International Workshop on Database and Expert Systems Application*. Zaragoza, 2004. p. 444-448. Disponível na Internet: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1333514>.

WAZLAWICK, R. S. (2008) - *Metodologia de pesquisa para Ciência da Computação*. Rio de Janeiro : Elsevier, 184 p. ISBN 978-85-352-3522-7.