Revista Digital de Biblioteconomia e Ciência da Informação www.sbu.unicamp.br/seer/ojs/index.php

ARTIGO

CDD: 025.04

CONSIDERAÇÕES ACERCA DO PROCESSO DE ALIMENTAÇÃO DE REPOSITÓRIOS ATRAVÉS DA IMPORTAÇÃO DE REGISTROS DE BASES DE DADOS INTERNACIONAIS[1]

CONSIDERATIONS AROUND OF PROCESS OF UPLOADING REPOSITORIES VIA IMPORT RECORDS OF INTERNATIONAL DATA BASES

Renan Carvalho Ramos[2]
Pedro Ivo Silveira Andretta[3]
Eduardo Graziosi Silva[4]

Resumo

Considerando que há um interesse global e crescente na implementação de repositórios institucionais e que as discussões em âmbito nacional sobre esse tema tem sido modestas, e que poucas são as iniciativas consolidadas no desenvolvimento desses sistemas nas universidades brasileiras, este trabalho toma como tema a questão da carga automática de registros comentada pelas diretrizes do programa intercontinental Alfa Biblioteca Babel. Para demonstrar as possibilidades e limitações da carga automática de registros referenciais a partir de bases de dados internacionais foi feito um estudo que se dividiu em quatro etapas, que se resumem em: escolha de um software para simular um repositório; análise de formatos de exportação dos metadados de algumas bases de dados disponíveis no Portal de Periódicos Capes; prospecção de registros e conversões de formato dos metadados; e alimentação do repositório através da importação dos registros coletados, indicando os efeitos desta prática. Como resultados são apontados um conjunto de ponderações a respeito do método que deve ser empregado para um melhor uso da carga automática de registros. Acredita-se que a carga automática de registros referenciais tem muito a colaborar com o início do desenvolvimento dos repositórios, à medida que oferece visibilidade à produção cientifica, favorecendo as instituições na elaboração dos índices de produção científica extraídos a partir dos emergentes estudos webmétricos, sem desrepeitar os principios de direitos do autor ou das editoras.

Palavras-chave: Bases de dados internacionais. Carga automática de registros. Importação de registros. Repositórios institucionais. *Software GreenStone*.

Abstract

Whereas there is a growing global interest in the implementation of institutional repositories and discussions in the national sphere about this issue have been modest, and there are few consolidated initiatives to develop these systems in Brazilian universities, this research takes as its theme the issue of automatic loading of records commented by the guidelines of the intercontinental program Alfa Biblioteca Babel. In order to demonstrate the possibilities and limitations of automatic loading of reference records from international databases a four-stage study was done, which summarizes in: choosing of a software to simulate a repository; analysing export formats of the metadata of some databases available in the Portal Periódicos Capes; prospecting records and metadata format conversions; and feeding the repository by means of the importation of collected records, indicating the effects of this practice. As a

[1] Este texto reproduz as discussões apresentadas pelos autores no artigo "A implementação de repositórios a partir da importação de registros de bases de dados internacionais: possibilidades e limitações" apresentado no XXIV Congresso Brasileiro de Biblioteconomia, Documentação e Ciência da Informação.

[2] Graduação em Biblioteconomia e Ciência da Informação pela Universidade Federal de São Carlos (2009). Bibliotecário da Universidade Estadual Paulista Júlio de Mesquita Filho. E-mail: renan@nit.ufscar.br.

[3] Universidade Federal de São Carlos. E-mail: andretta_pedro@yahoo.com.br – São Carlos, SP, Brasil.

[4] Biblioteca Camargo Silva, Dias de Souza Advogados. E-mail: eduardograziosi@ig.com.br – Sorocaba, SP - Brasil

result, a number of cogitations were pointed, regarding the method that should be employed for a better use of the automatic loading of records. It is believed that the automatic loading of reference records can collaborate a lot with the early development of repositories, as it offers visibility to the scientific production, favoring institutions in the production of rates of scientific production extracted from emerging webmetrics studies, without disrespecting the principles of copyright of authors or publishers.

Keywords: *International databases. Automatic loading of records. Records import. Institutional repositories. GreenStone Software.*

INTRODUÇÃO

As instituições de ensino superior brasileiras são responsáveis por grande parte da produção científica do país. Diante da necessidade de promover o acesso ao que já foi publicado e de disponibilizar para a sociedade, torna-se importante o desenvolvimento de repositórios institucionais "[...] como uma forma de minimizar os problemas de acesso aos documentos, permitindo reunir, preservar e divulgar, por meio de arquivos digitais, a produção científica de uma instituição, proporcionando maior transparência aos investimentos em ciência no país." (PAVÃO; SOUSA; CAREGNATO, 2009).

Considerando, como afirma Carvalho (2009), o crescente interesse pela implementação de repositórios institucionais por todo o mundo e que as discussões em âmbito nacional sobre esse tema tem sido modestas nas universidades brasileiras, além de serem poucas as iniciativas consolidadas no desenvolvimento destes sistemas, este trabalho toma como tema a questão da carga automática de registros, comentada pelas diretrizes do programa intercontinental Alfa Biblioteca Babel (REDE..., ([200?]).

É objetivo desta pesquisa demonstrar a potencialidade da carga automática de registros referenciais contidos em grandes bases de dados internacionais, assim como elaborar um plano geral para este procedimento, oferecendo subsídios teóricos e técnicos para a implementação e carga inicial de registros em repositórios institucionais. Além disso, são descritas algumas possíveis limitações deste procedimento.

REVISÃO DE LITERATURA

Desde o ano de 1991 têm surgido iniciativas em prol do compartilhamento do saber cientifico com os primeiros repositórios de *e-prints*. Em 2001, com a conferência *Budapest Open Access Initiative*, foi recomendada a proposta de acesso aberto à informação científica, através de

duas estratégias: a via dourada e via verde. A via dourada recomenda a publicação de artigos em periódicos científicos de acesso aberto, isto é, sem restrições de acesso ou uso; e a via verde, trata da permissão dos editores científicos para que os autores de trabalhos publicados armazenassem sua produção em repositórios, em especial repositórios institucionais. Nesse contexto, a Declaração de Berlim (2003) aponta que o estabelecimento do acesso aberto requer o empenho de todos os envolvidos na produção do conhecimento científico. Para tanto, deve satisfazer duas condições: os autores e detentores dos direitos devem permitir que os usuários acessem seus conteúdos gratuitamente, podendo copiá-los, distribuí-los, transmiti-los e exibi-los publicamente, em qualquer suporte e desde que atribua corretamente a autoria; e deve ser depositada uma versão completa da obra e de seus materiais complementares, em formato eletrônico, em um repositório que utilize normas técnicas adequadas e que seja mantido por uma instituição.

Os repositórios são sistemas de informação que possuem ferramentas que permitem importar, armazenar, preservar, recuperar e exportar objetos digitais. Leite (2009) apresenta para esses sistemas tipologias segundo seu escopo, a saber: Institucionais, Temáticos ou de Teses e Dissertações.

Neste contexto, as bibliotecas universitárias, possuidoras de experiência na gestão da informação em formas e suportes diversos, são as instituições que devem liderar a implementação dos repositórios nas universidades, contribuindo para a visibilidade da produção acadêmica e científica dessas, favorecendo a retroalimentação da pesquisa, a produção e suporte às publicações da universidade, bem como permitir o acesso às informações por elas produzidas (REDE..., ([200?], tradução nossa).

As missões dos repositórios podem ser resumidas em: ampliar a comunicação científica, permitindo tanto o livre acesso aos arquivos como também a indexação e recuperação destes conteúdos através de buscadores; e a preservação do patrimônio intelectual de um autor, cientista, comunidade ou instituição. Além disso, permitem o autoarquivamento, isto é, depósito de conteúdo pelos próprios autores, de materiais publicados ou não, que "[...] podem ser recarregados no depósito institucional e/ou em outros repositórios temáticos." (REDE..., ([200?]), p. 37, tradução nossa), justificando, assim, a adoção da via verde pelos editores científicos.

Segundo podemos notar na nascente literatura que vem se formando sobre o tema "repositórios", os problemas de implantação destes são: por um lado de ordem tecnológica e técnica, que vem sendo resolvidas por esforços do IBICT, a exemplo do estudo de Leite (2009); e por outro as barreiras dos direitos autorais. Visando solucioná-los, Rede ([200?], tradução nossa)

apresenta as seguintes etapas para implementação dos repositórios:

1. Seleção do *software* para gestão dos conteúdos, que permita customização segundo sua demanda;

2. Criação de mecanismos de controle de qualidade visando agregar credibilidade ao repositório;

3. Padronização dos metadados que descrevem e identificam os conteúdos, como o formato MARC e o XML, respectivamente;

4. Estabelecimento de políticas acerca da propriedade intelectual, que deve considerar aspectos como o tipo de documento a ser incluído, seus direitos legais e a autorização para publicá-los.

Esses documentos podem ser, segundo Rede ([200?], tradução nossa), os produtos científicos (teses, patentes, *software*, etc.), produtos institucionais e/ou administrativos (regulamentos, relatórios técnicos, dentre outros) e os objetos de aprendizagem (anotações da aula, *blogs*, etc.). A gestão dos conteúdos deve seguir a Política de Repositório de Objetos de Aprendizagem e do Conhecimento estabelecida. Assim, o autoarquivamento pode ser realizado cumprindo-se as seguintes etapas: fase de identificação ou validação informativa e identificação do agente (informações que identifiquem o autor); fase de informação legislativa e institucional (concordância do autor com a política do repositório); fase de recarga; e fase de avaliação do documento (REDE..., ([200?], tradução nossa).

Sobre a fase de recarga, tema desta pesquisa, Rede ([200?]), p. 42, tradução nossa) descreve-a como sendo

> A prática mais usual nos dias de hoje é o auto-arquivo da produção, feita pelo próprio autor (docente ou investigador), no entanto, o lançamento do repositório deveria estar precedido de uma carga inicial importante que lhe outorgasse suficiente opinião crítica para dar credibilidade ao projeto, possibilitar que o auto-arquivo seja aceito pela comunidade e ganhe aderentes.
> Cada instituição deverá verificar se nessa carga inicial, dispõe dos metadados potenciais, decorrentes de alguma base de dados de gestão interna, ou de bases de dados comerciais tais como *Scopus* ou *ISI Web of Knowledge*. Neste caso deverá avaliar a possibilidade de fazer uma recarga automática com posterior revisão de registros, ao invés de uma recarga manual.

Considerando que a carga automática confere credibilidade ao repositório, isso pode favorecer os colaboradores a depositarem suas produções acadêmicas e científicas nos

repositórios. Contudo, as instituições mantenedoras desses sistemas devem desenvolver uma política de desenvolvimento de coleção e divulgá-la para os autores da instituição, a fim de fomentar seu uso e crescimento pela comunidade.

MATERIAIS E MÉTODOS

Para demonstrar as possibilidades e limitações da (re)carga automática de registros referenciais a partir de bases de dados internacionais realizou-se um estudo que se dividiu em quatro etapas.

Inicialmente averiguaram-se os *softwares* disponíveis atualmente para a construção de repositórios, considerando os estudos avaliativos de Tramullas e Garrido (2006). A escolha deste artigo para balizar a eleição do *software* considera, sobretudo, sua extensa bibliografia que reuniu parte importante, ou ainda muito citada, dos trabalhos relacionados à avaliação de ferramentas para o desenvolvimento de repositórios. A partir desta publicação foram buscados, ainda, os artigos de Crow (2002), Wang, Assion, Matthaei (2003) e Han (2004).

Na segunda etapa, quando já determinado o *software* mais adequado para a pesquisa, foram explorados quais esquemas de metadados são aceitos para dar início a uma exportação de registros referenciais a partir das bases de dados assinadas pelo Portal de Periódicos Capes[5]. Em seguida, procedeu-se à visita de determinadas bases de dados, cujos formatos de saída foram tabulados (Quadro 1), explorando-se também a capacidade de algumas ferramentas capazes de converter determinados esquemas de metadados.

Para a terceira etapa foi eleita uma instituição de ensino superior assim como determinado um recorte temporal para a coleta de dados e criada uma estratégia de busca capaz de recuperar grande parte dos registros pertencentes à essa instituição em três bases de dados, a saber: Compendex[6], Scopus[7] e Web of Science[8], dando inicio à prospecção dos registros.

Por fim, na quarta etapa, iniciou-se a alimentação do repositório através da importação de metadados em esquema Bibtex, verificando as possibilidades e limitações deste processo.

[5] <http://novo.periodicos.capes.gov.br/>.
[6] <http://www.engineeringvillage2.org>
[7] <http://www.scopus.com>.
[8] < http://apps.isiknowledge.com>

RESULTADOS FINAIS

A partir dos estudos realizados sobre os *softwares* disponiveis para o desenvolvimento de um respositório foi escolhido, para este experimento, o Greenstone, uma ferramenta *open source*, criada e mantida pela *University of Waikato* e distribuído gratuitamente pela UNESCO e HumanInfo NGO. O Greenstone admite em sua coleção documentos digitais em diversos formatos, como por exemplo: .doc, .pdf, .xls, .ppt, XML, HTML, além de conjuntos de metadados em esquemas Refer, BibTex, Dublin Core entre outros, permitindo também a exportação de seus registros para sistemas como DSpace, Fedora ou ainda qualquer outro que use metadados METS ou MARCXML.

Após a escolha do *software* utilizado para simular um repositório procedeu-se a um levantamento para identificar os formatos de saída dos registros bibliográficos em algumas bases de dados disponíveis no Portal de Periódicos Capes, obtendo a seguinte relação:

QUADRO 1
Esquemas de metadados[9]

Base de dados	Esquemas									
	Bib TeX	HTML	Texto	RIS	RefWorks	Csv	XML	MARC21	EndNote	OpenURL COinS
Web of Science	x	x	x							x
Scopus	x			x	x	x				x
Compendex	x		x	x	x					x
EBSCO	x	x	x	x	x		x	x	x	x
CSA Illumina		x	x		x					x
Wilson Web			x	x	x					

A instituição de ensino e pesquisa eleita para servir de teste foi a Universidade Federal de São Carlos com um recorte temporal compreendendo único e exclusivamente as publicações do ano de 2001[10]. Desta forma, foram criadas as seguintes estratégias de busca:

[9] FONTE – *Web of Science, Scopus, Compendex, EBSCO, CSA Ilumina, Wilson Web*.
[10] Devido ao fato desta pesquisa deter um caráter experimental, foi desconsiderada a cobertura total haja vista que os resultados obtidos com essa amostragem podem ser generalizáveis.

QUADRO 2
Estratégias de busca[11]

Base de Dados	Expressão Booleana	Quant. de registros
Compendex	((ufscar OR universidade federal sao carlos OR univ* federal sao carlos OR univ* fed* sao carlos OR federal university of sao carlos) WN AF)	212
Scopus	((AFFIL(ufscar OR universidade federal sao carlos OR univ* federal sao carlos OR univ* fed* sao carlos)) OR (AFFIL(federal university of sao carlos))) AND (PUBYEAR IS 2001)	229
Web of Science	Address=(ufscar OR universidade federal sao carlos OR univ* federal sao carlos OR univ* fed* sao carlos OR federal university of sao carlos) AND Year Published=(2001)	423

Como anteriormente visto, as bases de dados Web of Science, Scopus e Compendex permitem a prospecção de registros em diversos formatos, inclusive BibTex, esquema esse, aceito pelo Greenstone. Considerando isso, foram coletados registros em formato não aceitos pelo Greenstone, tal como RIS e OpenURL CoinS, que foram convertidos para BibTex com o propósito de averiguar o efeito desta operação, comparando-se a estrutura dos metadados. Essa experimentação se justificativa na medida que a conversão dos metadados poderia ser usada na prospecção de registros de bases de dados menos robustas, como, por exemplo, Scielo[12], Scirus[13] (7), Pubmed[14] (8) e Redalyc[15] (9)..

Para converter os metadados foi adotado o gestor de referências Zotero, uma ferramenta *open source* desenvolvida e distribuída gratuitamente pela *Center for History and New Media da George Mason University (GMU)*. O Zotero tem ganhado destaque junto à comunidade científica, tendo, entre outras funções, a exportação de sua "biblioteca" para extensões Zotero RDF, MODS, Refer/BibIX, RIS, Unqualified Dublin Core RDF, Wikipedia Citation Templates e BibTeX.

Conforme analisado, a conversão de esquema de metadados para BibTeX, usando o sistema de gerenciamento de referências Zotero, revelou-se bastante eficiente, contudo, observou-se a ausência do metadado "document_type=", presente quando realizada a prospecção de registros completos diretamente das bases de dados. Desse modo, apesar do *software* Greenstone

[11] FONTE – *Compendex, Scopus e Web of Science*.
[12] < http://www.scielo.org/php/index.php>.
[13] <http://www.scirus.com/>.
[14] <http://www.ncbi.nlm.nih.gov/pubmed>.
[15] <http://redalyc.uaemex.mx/>.

distinguir coleções de "articles", "proceedings", "books" gerados de acordo com os metadados, verificou-se que os registros são descritos sempre como "article", recomendando-se, assim, limitar a busca, a princípio, apenas por documentos que sejam artigos de periódicos, até que sejam desenvolvidas adequações.

Tendo em mãos os conjuntos de registros com a extensão BibTex, foi realizada a importação desses no *software* Greenstone em sua versão 2.81 para Windows através da Interface do bibliotecário (*Greenstone Librarian Interface – GLI*), adotada por parecer mais intuitiva. Uma vez importados os registros, desenhou-se o formato de exibição dos registros de forma bastante simples assim com a interface do usuário, com o propósito de analisar os registros.

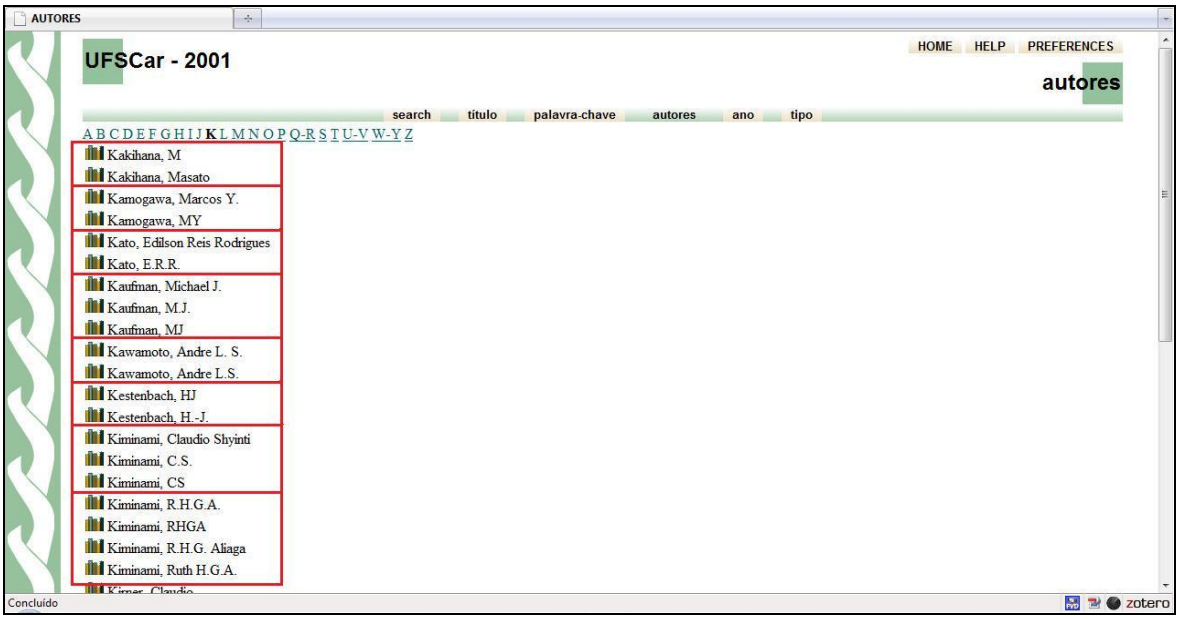

FIGURA 1 – Interface do Usuário – Organização por Autor[16]

A dispersão da coleção de registros de autor (ver contorno em vermelho), apresentada acima, deve-se à variedade de formas como um mesmo autor aparece indexado nas diversas bases de dados (ora com nome completo, ora abreviado, ora com ponto depois das iniciais, ora sem ponto). Destaca-se, portanto, a ausência de padronização dos dados de autoria no momento de sua entrada nas bases de dados analisadas, interferindo negativamente na recuperação da informação pelos usuários. Vale recordar que problema semelhante pode ser percebido, por exemplo, no repositório da UNICAMP, UNIFESP, UNESP, USP, ITA e UFSCar que trabalham com os dados

[16] FONTE – Os autores.

do Scielo (10).

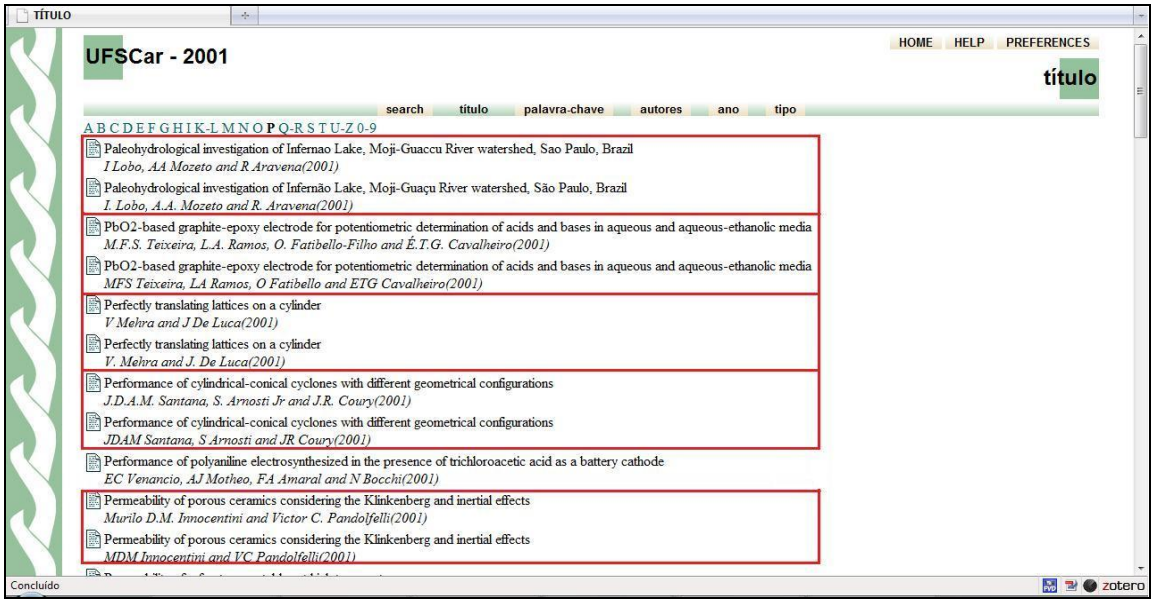

FIGURA 2 – Interface do Usuário – Organização por Título

No que se refere à disposição dos registros por ordem de título, observou-se a duplicação de registros (com o contorno vermelho) pelo fato de que algumas das bases selecionadas indexam os mesmos periódicos. Sobre isso, Torres-Salinas e Jiménes-Contreras (2009, p. 203, tradução nossa) comentam que "de todas das 8.199 revistas incluídas no JCR [Thomson Reuters], identificaram-se 7.825 na fonte de Scopus, ou seja, há uma sobreposição de 95%"

Para evitar a duplicação de registros, recomenda-se que ao realizar a busca de registros em cada base de dados sejam identificados os periódicos exclusivos de cada uma e quando ocorrer de duas ou mais bases possuírem o mesmo periódico, recomenda-se utilizar aquela com maior número de registros para determinado periódico. Nesse sentido, podemos observar que tem sido frequente nas bases de dados a opção de refinamento da busca, não sendo necessário criar estratégia de busca para delimitar a cobertura dos periódicos. Ressalta-se, ainda, que quanto maior o número de bases consultadas maiores serão as possibilidades de se conseguir uma cobertura completa dos trabalhos publicados por um grupo de pesquisadores ou instituição.

CONSIDERAÇÕES FINAIS

A prática da carga automática de registros referenciais pode colaborar com o início do desenvolvimento dos repositórios, à medida que oferece visibilidade à produção científica, sem entrar em conflito com os direitos autorais, além de favorecer as instituições nos índices de produção científica extraídos a partir dos emergentes estudos webmétricos.

Esta pesquisa demonstrou que é possível a carga automática de registros referenciais a partir da importação e conversão de metados originários de bases de dados internacionais, sendo necessário, contudo, realizar tratamentos com vista à padronização das entradas de "autor" e tomados alguns cuidados para evitar a duplicação de "títulos" na coleção.

Como perspectiva, ressalta-se a necessidade de estudos voltados ao tratamento descritivo dos dados e o comportamento de determinados metadados durante a importação e exportação de registros tais como "Tipo do documento", "Palavras-chave" e "Abstracts".

As etapas utilizadas nesta pesquisa consistem em recomendações que podem ser aplicadas para os demais tipos de repositórios apresentados anteriormente por Leite (2009), a exemplo do repositório temático, no qual seria feito uma estratégia de busca nas bases de dados que contemplasse os termos ligados à área abrangida pelo mesmo.

REFERÊNCIAS

CARVALHO, M. C. R. de. Bibliotecas universitárias brasileiras e a implantação de repositórios institucionais. **Revista Informação & Universidade**, Rio de Janeiro, v.1, n.0, p. 1-9, jul./dez., 2009. Disponível em: <http://www.siglinux.nce.ufrj.br/~gtbib/site/2009/06/implantacao-de-repositorios/>. Acesso em: 07 set. 2009.

CROW, R. **A guide to institutional repository software**. New York: Open Society Institute, 2004.

DECLARAÇÃO DE BERLIM SOBRE ACESSO LIVRE AO CONHECIMENTO NAS CIÊNCIAS E HUMANIDADES. [S.l.: s.n.], 2003. Disponível em: <http://oa.mpg.de/openaccess-berlin/BerlinDeclaration_pt.pdf>. Acesso em: 28 jun. 2010.

HAN, Y. Digital content management: the search for a content management system. **Library Hi Tech**, v. 22, n. 4, p. 355-365, 2004. DOI 10.1108/07378830410570467.

LEITE, F. C. L. **Como gerenciar e ampliar a visibilidade da informação científica brasileira**: repositórios institucionais de acesso aberto. Brasília: IBICT, 2009. Disponível em: <http://www.ibict.br/anexos_noticias/repositorios.institucionais.F.Leite_atualizado.pdf>. Acesso em: 28 dez. 2009.

PAVÃO, C. G.; SOUSA, R. S. C. de; CAREGNATO, S. E. Publicização da literatura científica através de repositórios institucionais. In: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA, DOCUMENTAÇÃO E CIÊNCIA DA INFORMAÇÃO, 23., 2009, Bonito. **Anais... .**Bonito: FEBAB, 2009. CD-ROM.

REDE ALFA BIBLIOTECA DE BABEL. **Diretrizes para a criação dos repositórios institucionais nas universidades e organizações de educação superior**.Valparaíso: [s.l.], ([200?]). Disponível em: <http://www.sisbi.uba.ar/institucional/proyectos/internacionales/Directrizes_RI_portugues.pdf> Acesso em: 02 jan. 2010.

TORRES-SALINAS; JIMÉNES-CONTRERAS. Introducción y estudio comparativo de los nuevos indicadores de citación sobre revistas científicas en Journal Citation Reports y Scopus. **El profisional de la información**, Barcelona, v. 19, n. 2, mar-abr., 2009.

TRAMULLAS SAZ, J.; GARRIDO PICAZO, P. Software libre para repositorios institucionales: propuesta para un modelo de evaluación de prestaciones. **El profisional de la información**, Barcelona, v. 15, n. 3, maio-jun., 2006.

VIANA, C. L. M.; MÁRDERO ARELLANO, M. A.; SHINTAKU, M. Repositórios institucionais em ciência e tecnología: uma experiencia de customização do DSpace. In: FUJITA, M. S. L. (Org.). **A dimensão social da biblioteca digital na organização e acesso ao conhecimento**: aspectos teóricos e aplicados. São Paulo, SP: SIBI/USP, 2005. (v. 1). 805 p.

WANG, J. Y.; ASSION, M.; MATTHAEI, B. **Inventories–open archives software tools**. Open Archives Forum, 2003. Disponivel em: <http://www.oaforum.org/otherfiles/tv–tools.pdf>. Acesso em: 23 jun. 2010.

Recebido em: 29/08/2010
Publicado em: 13/07/2012