Pesquisa científica

F. Ciências Sociais Aplicadas - 11. Documentação e Informação Científica - 1. Documentação e Informação Científica

DA POSSIBILIDADE DE UMA *WEB OF SCIENCE* PARA A AMÉRICA LATINA E CARIBE: EXTRAÇÃO AUTOMÁTICA DE UMA BASE DE CITAÇÕES DO SCIELO PARA O PERIÓDICO PERSPECTIVAS EM CIÊNCIA DA INFORMAÇÃO E PARA A COLEÇÃO DE SAÚDE PÚBLICA

MAX CIRINO DE MATTOS - DOUTORANDO - ESCOLA DE CIÊNCIA DA INFORMAÇÃO – UFMG
BEATRIZ VALADARES CENDÓN - PROFA. DRA./ORIENTADORA – ESCOLA DE CIÊNCIA DA INFORMAÇÃO - UFMG

INTRODUÇÃO:

Para Guimarães et al. (2011, p. 5), a deficiência da cobertura dos índices existentes (a exemplo dos produzidos pelo ISI) pode ocasionar deformações "nos processos de gestão das atividades científicas em contexto local", o que levou vários países a buscar o desenvolvimento de índices de citações locais: China (XIN-NING, 2001), Polônia (WEBSTER, 1998) e União Europeia (GOGOLIN et al., 2003) são exemplos citados.
Meneghini (1998) apresenta uma análise da produção científica brasileira e ressalta a carência de bases de dados em informação científica nos moldes do SCI.
Conforme Cendón *et al.*, (2012), "a produção de indicadores que possam medir e avaliar a produção científica brasileira passa necessariamente pela existência de um índice de citações, nos moldes daqueles produzidos pelo *Institute of Scientific Information* (ISI)". Os autores afirmam que tais indicadores "podem ser balizadores de políticas científicas nacionais, entre várias outras aplicações".
O processo proposto busca a obtenção automática dos metadados dos artigos e referências citadas disponíveis no formato XML para a criação de uma base de citações para América Latina e Caribe, extensivo a outras coleções, considerando sua fonte primária: o *Scientific Electronic Library On-line* (SCIELO).

OBJETIVO DO TRABALHO:

Desenvolver um processo para obtenção e interpretação dos arquivos XML do SCIELO, gerando uma base de citações atualizada contínua e automaticamente para as coleções de periódicos disponíveis.

MÉTODOS:

Inicialmente foram identificados os padrões de composição do endereço eletrônico de cada coleção do SCIELO (África do Sul, Brasil, Chile, Colômbia, Cuba, Espanha, México, Portugal, Saúde Pública e Venezuela) para a obtenção da lista de periódicos, dos dados fonte anuais (total de fascículos, artigos e citações de cada periódico) e dos arquivos XML.
Após a identificação automática dos periódicos de cada uma dessas coleções foram obtidos os dados fonte anuais disponíveis no SCIELO para cada periódico.
Para o teste inicial foram obtidos os arquivos XML do periódico Perspectivas em Ciência da Informação (PCI).
Após a validação do teste inicial, foram obtidos os arquivos XML dos periódicos da Coleção de Saúde Pública. Esses arquivos foram interpretados e os metadados de cada artigo e de cada uma de suas referências foram arquivados.
Todas essas informações foram gravadas em um banco de dados MySQL a partir de programas gerados com a linguagem PHP.
Os dados fonte obtidos do SCIELO foram comparados às informações obtidas a partir dos arquivos XML.

RESULTADOS E DISCUSSÃO:

Para o teste inicial com o periódico PCI foram identificados nos dados do SCIELO 24 fascículos, 290 artigos e 7.714 citações. O resultado obtido a partir do processamento dos arquivos XML apresentou os mesmos números de fascículos e citações, com 292 artigos.
Os dados obtidos do SCIELO para as coleções identificadas totalizaram 880 periódicos (sendo 765 correntes e 115 não correntes), 26.412 fascículos, 330.623 artigos, 8.662.607 citações concedidas e 657.362 citações recebidas.
A coleção do Brasil destacou-se com 305 periódicos, 13.396 fascículos (50,72%), 196.088 artigos (59,31%) e 5.165.748 citações (59,63%).
A coleção investigada - Saúde Pública - apresentou 14 periódicos com dados fonte disponíveis, 1.372 fascículos, 17.899 artigos e 494.083 citações. Para esta coleção, os dados obtidos dos arquivos XML

apresentaram informações divergentes: 1.329 fascículos, 18.656 artigos, 490.865 citações. A diferença total de citações identificadas não foi significativa: 3.218 (0,65% do total do SCIELO).
Do total de 205 dados fonte anuais avaliados, em 157 (76,58%) as citações encontradas nos arquivos XML coincidiram com as informações do SCIELO; em 39 (19,02%) registros os dados do SCIELO foram maiores e em 9 (4,40%), menores.

CONCLUSÕES:

O elevado índice de acerto na identificação das citações a partir dos arquivos XML comprovou a eficácia do processo desenvolvido para a criação da base de citações de coleções do SCIELO.
Em alguns casos foi possível a identificação de 4 tipos de problemas: estrutura de *tags* do XML incompleta, impedindo a importação das citações; arquivos XML não acessíveis, com erro de estrutura XML; dados fonte de periódicos que estavam disponíveis no início do projeto mas foram excluídos do SCIELO (nesses casos, os arquivos XML incorporados geravam diferenças) e diferentes dados fonte, em coleções diferentes, para um mesmo periódico, gerando problemas na obtenção dos arquivos XML.
Mais detalhes sobre a forma como o SCIELO calcula os dados sobre número de fascículos, artigos e citações precisam ser investigados para a análise das diferenças encontradas.
A base de citações gerada para a Perspectivas em Ciência da Informação será disponibilizada em breve no site desse periódico.
Nova pesquisa está em andamento, buscando a obtenção de todos os arquivos XML das coleções citadas para a construção de uma base de citações para cobertura da América Latina, Caribe e outras coleções do SCIELO.

Palavras-chave: Índice de citações nacionais
Controle da produção científica nacional
Indicadores científicos nacionais