Criação automática de uma base de citações para o SciELO a partir dos seus arquivos XML

Max Cirino de Mattos
Universidade Federal de Minas Gerais (UFMG), Brasil. E-mail: max@cognotec.com.br

Beatriz Valadares Cendón
Universidade Federal de Minas Gerais (UFMG), Brasil. E-mail: bcendon@gmail.com

Resumo

O artigo demonstra o uso arquivos *eXtensible Markup Language* (XML) da *Scientific Electronic Library Online* (SciELO) para criação de um protótipo para a criação de bases de citações das suas revistas. O foco principal do artigo é a descrição da metodologia usada para a obtenção automática dos dados estatísticos de cada Coleção do SciELO, bem como dos arquivos XML disponíveis para cada periódico. Esses arquivos foram interpretados e os metadados dos artigos e das referências usadas na sua produção foram gravados automaticamente em uma base de citações. A Coleção Saúde Pública foi usada para exemplificar a aplicação do protótipo. Sugere-se a disponibilização da base de citações, com atualização automática, de forma integrada ao site de cada uma das revistas listadas.
Palavras-chave: Ciência da Informação. SciELO. Base de Citações. XML.

1 Introdução

De acordo com Cendón, Guimarães, Silva, Oliveira, Mattos, Santana e Fernandes (2012, p.2), "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".

Guimarães, Silva, Santana, Braga, Bochner e Goldbaum (2011, p. 5) explicam que 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". Conforme os autores, essas deformações podem levar diversos países a buscar o desenvolvimento de índices de citações locais, a exemplo da China (XIN-NING, 2001), Polônia (WEBSTER, 1998) e União Europeia (GOGOLIN et al., 2003).

A carência de bases de dados em informação científica nos moldes do *Science Citation Index* (SCI) também é ressaltada por Meneghini (1998) em sua análise da produção científica nacional.

Visando suprir essa carência, a pesquisa desenvolvida na Escola de Ciência da Informação da UFMG buscou identificar a viabilidade da criação de bases de citações considerando como fonte primária a *Scientific Electronic Library On-Line* (SciELO) a partir da obtenção automática dos metadados dos artigos e referências citadas disponíveis no formato *eXtensible Markup Language* (XML). Este artigo relata a metodologia desenvolvida para obtenção dos dados e para a interpretação dos mesmos, e discute possíveis aplicações, usando como exemplo a Coleção Saúde Pública do SciELO.

2 Metodologia

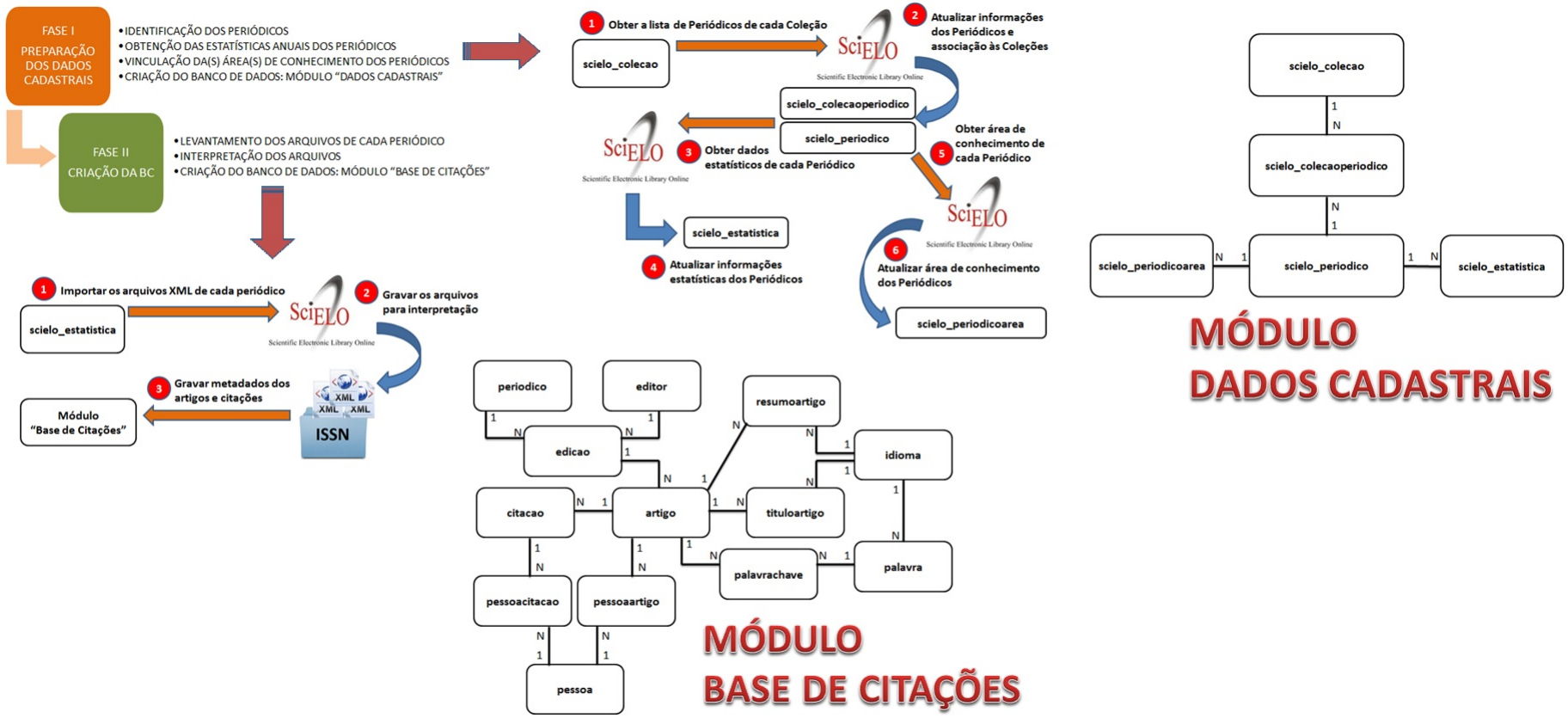
Este trabalho classifica-se como uma pesquisa aplicada (LAKATOS; MARCONI, 2007, p.20) ou exploratória, segundo Tripodi *et al.* (1975), que definem o propósito da pesquisa exploratória como a demonstração da viabilidade de um determinado programa ou técnica como uma solução em potencial para problemas práticos – objetivo central da criação da base de citações do SciELO.

Também pode ser classificada como híbrida, ao utilizar-se de métodos qualitativos e quantitativos (CRESWELL; CLARCK, 2011). Nesse sentido, trata-se de um trabalho qualitativo enquanto processo de desenvolvimento de um modelo relacionado à abstração, generalização e formação de conceitos para a metodologia proposta; e quantitativo, na medida em que o desenvolvimento do protótipo e da base de citações relacionam-se ao estudo objetivo da bibliometria. A combinação desses dois tipos pode proporcionar uma base contextual mais rica para interpretação e validação dos resultados, e de acordo com a definição dos autores esta pesquisa é interativa de prioridade qualitativa: interativa, pois os métodos qualitativo e quantitativo interagem e são integrados antes da análise final; a prioridade qualitativa refere-se ao propósito central de criação e abstração de um modelo para representação do processo de automação da construção da base de citações, e possui maior ênfase que a parte quantitativa, sendo esta última mais simples de acordo com a abordagem proposta.

A técnica usada na consecução dos passos metodológicos apresentados adiante tem como base – apesar de não seguir rigidamente suas notações – a modelagem relacional proposta por Codd (1969; 1970) e o Modelo Entidade-Relacionamento (CHEN, 1976; 2002). O uso desta técnica visa, em parte, a generalização do modelo construído para aplicação em outras coleções digitais.

A visão geral da metodologia completa é consolidada na Figura 1 a seguir, que apresenta os dois passos e suas etapas, e o modelo relacional criado para os Módulos “Dados Cadastrais” e “Base de citações”, que são detalhados adiante, respectivamente nos tópicos 2.2.4 e 2.3:

Figura 1 - Visão geral da metodologia: passos e modelagem de dados simplificada



Fonte: desenvolvida pelos autores

A seguir são descritos os passos metodológicos utilizados para o desenvolvimento deste trabalho, desde a escolha as ferramentas tecnológicas, passando pelo desenvolvimento dos experimentos e criação de sistemas.

2.1 Definição das ferramentas para desenvolvimento

Para o desenvolvimento do protótipo e realização dos experimentos descritos foi usado o MySQL, um sistema de gerenciamento de banco de dados (SGBD) com base na *General Public License* (GPL), e que apresenta uma fácil integração com a linguagem de programação PHP, além de ser multiplataforma (funciona tanto no sistema operacional Windows como no sistema operacional Linux), e ter excelente desempenho e estabilidade (GUIMARÃES; SILVA; SANTANA; BRAGA; BOCHNER; GOLDBAUM, 2011).

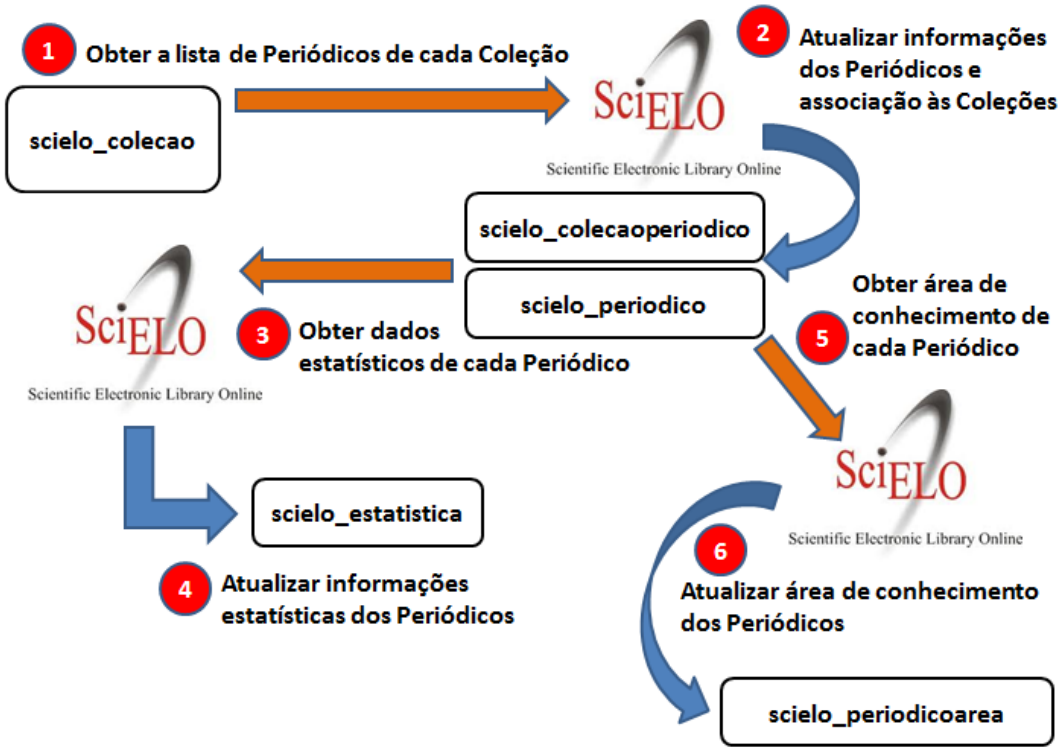
O ambiente de desenvolvimento dos experimentos apresenta a seguinte configuração: sistema operacional Windows 7 *Home Premium*, *service pack* 1, 64 bits; editor de PHP Zend Studio 5.0.0 e Zend Guard 4.0.0 para criptografia dos programas a serem disponibilizados na internet; SQLyog 7.02 para manipulação do banco de dados MySQL. No ambiente *web* funcionam os programas PHP desenvolvidos, criptografados com a ferramenta *Zend Guard* e transmitidos com a ferramenta FileZilla; o banco de dados MySQL é administrado a partir do uso da ferramenta PHPMyAdmin em ambiente Linux.

Os navegadores Chrome e Firefox foram usados ao longo do desenvolvimento dos experimentos e desenvolvimento de sistemas, sempre atualizados com a versão mais recente. Os dois navegadores foram usados aleatoriamente, tanto no ambiente de desenvolvimento quanto no ambiente *web*.

2.2 FASE I: Obtenção dos dados do SciELO

A preparação dos dados cadastrais, conforme a Figura 2 adiante, é realizada a partir da obtenção de dados do SciELO, e pode ser detalhada em seis passos: (1) obtenção da lista de periódicos a partir da interpretação da página HTML (cujo endereço está endereço gravado na tabela *scielo_colecao*) de cada coleção; (2) atualização das informações sobre cada periódico pertencente às coleções (os dados são gravados em *scielo_periodico* e *scielo_colecaoperiodico*); (3) obtenção dos dados fonte (número de fascículos, artigos e citações em cada ano) para cada periódico e (4) armazenamento dessas informações na tabela *scielo_estatistica*; (5) obtenção da(s) área(s) de conhecimento de cada periódico e (6) gravação as informações em *scielo_periodicoarea*.

Figura 2 - FASE I: Preparação dos dados cadastrais



Fonte: Mattos(2013, p. 19).

Esses passos foram agrupados em quatro etapas, descritas adiante, que tratam da obtenção dos dados do SciELO: 1) identificação dos periódicos; 2) obtenção dos dados estatísticos anuais de cada periódico; 3) vinculação da(s) área(s) de conhecimento; e 4) criação do Módulo "Dados Cadastrais" do banco de dados.

2.2.1 Identificação dos periódicos de uma Coleção do SciELO

Cada coleção do SciELO apresenta uma página com a lista alfabética completa de seus periódicos correntes e não correntes. No caso da Coleção de Saúde Pública, o *link* disponível em <http://www.scielosp.org/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso> apresenta esta relação, conforme a Figura 3:

Figura 3 - Relação de periódicos da Coleção Saúde Pública

Fonte: SciELO (2014[1]).

[1]Disponível em <http://www.scielosp.org/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso>. Acesso em:10 abril 2014.

Para a obtenção automática da relação de periódicos da Coleção foi desenvolvido um programa que interpreta o conteúdo deste *link* e extrai a situação (corrente ou não corrente), o título, o *International Standard Serial Number* (ISSN) e a quantidade de fascículos de cada periódico para armazenamento no banco de dados. Este programa analisa a referida página diariamente, mantendo atualizada a lista de periódicos.

Identificou-se nos links do SciELO para o acesso a múltiplas coleções, uma pequena variação no link de acesso à lista de periódicos, com o padrão a seguir:

http://DOMÍNIO/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso,

no qual o DOMÍNIO é apresentado no Quadro 1 a seguir para cada Coleção. É importante observar que a única exceção a este padrão foi a Coleção de Periódicos da Bolívia, que apresenta o final /**scielo.php/script_sci_alphabetic/lng_pt/nrm_iso**

Quadro 1 - Informações sobre as Coleções do SciELO: domínios

COLEÇÃO	**DOMÍNIO**	***LINK* da lista de periódicos**		
África do Sul	www.scielo.org.za	http://	www.scielo.org.za	/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso
Argentina	www.scielo.org.ar	http://	www.scielo.org.ar	/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso
Bolívia	www.scielo.org.bo	http://	www.scielo.org.bo	/scielo.php?script_sci_alphabetic/lng_pt/nrm_iso
Brasil	www.scielo.br	http://	www.scielo.br	/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso
Brasil *Proceedings*	www.proceedings.scielo.br	http://	www.proceedings.scielo.br	/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso
Chile	www.scielo.cl	http://	www.scielo.cl	/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso
Colômbia	www.scielo.org.co	http://	www.scielo.org.co	/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso
Costa Rica	www.scielo.sa.cr	http://	www.scielo.sa.cr	/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso
Cuba	scielo.sld.cu	http://	scielo.sld.cu	/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso
Espanha	scielo.isciii.es	http://	scielo.isciii.es	/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso
México	www.scielo.org.mx	http://	www.scielo.org.mx	/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso
Paraguai	scielo.iics.una.py	http://	scielo.iics.una.py	/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso
Peru	www.scielo.org.pe	http://	www.scielo.org.pe	/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso
Portugal	www.scielo.gpeari.mctes.pt	http://	www.scielo.gpeari.mctes.pt	/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso
Saúde Pública	www.scielosp.org	http://	www.scielosp.org	/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso
Social Sciences	socialsciences.scielo.org	http://	socialsciences.scielo.org	/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso
Uruguai	www.scielo.edu.uy	http://	www.scielo.edu.uy	/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso
Venezuela	www.scielo.org.ve	http://	www.scielo.org.ve	/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso
West Indian Medical Journal	caribbean.scielo.org	http://	caribbean.scielo.org	/scielo.php?script=sci_alphabetic&lng=pt&nrm=iso

Fonte: Mattos (2013, p.53)

2.2.2 Obtenção dos dados estatísticos anuais de cada periódico

Outras informações importantes apresentadas pelo SciELO estão contidas em uma lista de dados fonte para os periódicos indexados, a exemplo dos dados da Figura 4 a seguir[2]

Figura 4 - Lista de dados fonte: Revista Peruana de Medicina Experimental y Salud Pública

Data do último processamento 10-03-2014

Revista Peruana de Medicina Experimental y Salud Publica
ISSN 1726-4634

Lista de dados fonte

∇ - clique para selecionar a coluna de ordenação
Δ - indica a ordem corrente

titulo da revista/ano Δ	n. de fascículos ∇	n. de artigos ∇	n. de citações concedidas ∇	n. de citações recebidas ∇	média de artigos por fascículo ∇
Rev. perú. med. exp. salud publica	**16**	**399**	**1273**	**403**	**24.94**
2013	4	121	319	107	30.25
2012	4	100	254	83	25.00
2011	4	102	340	104	25.50
2010	4	76	360	109	19.00
total	**16**	**399**	**1273**	**403**	**24.94**

Fonte: SciELO (2014).

[2] Disponível em: <http://statbiblio.scielo.org//stat_biblio/index.php?state=15&lang=pt&country=spa&issn=1726-4634&CITED%5B%5D=REVISTA+PERUANA+DE+MEDICINA+EXPERIMENTAL+Y+SALUD+PUBLICA&YNG%5B%5D=all>. Acesso em: 10 abril 2014.

Para o acesso a essa página de dados fonte de cada periódico, foi identificado o padrão:

http://statbiblio.scielo.org//stat_biblio/index.php?state=15&lang=pt&country=scl&issn=ISSN&CITED%5B%5D=TITULO&YNG%5B%5D=all

Dessa forma, foi desenvolvido um programa para acessar, interpretar e armazenar os dados estatísticos anuais de cada periódico a partir de seu ISSN e título. Para o caso da PCI, foi usado o *link* com seu ISSN e título:

http://statbiblio.scielo.org//stat_biblio/index.php?state=15&lang=pt&country=scl&issn=1413-9936&CITED[]=PERSPECTIVAS+EM+CIENCIA+DA+INFORMACAO&YNG[]=all

Os dados obtidos sobre cada periódico são atualizados diariamente e gravados no banco de dados, conforme detalhamento no tópico descritivo do Módulo "Dados Cadastrais", adiante.

Para a importação automática das informações de Coleções do SciELO, levantou-se que a identificação das estatísticas de cada Coleção também segue um padrão determinado, que inclui, além do ISSN e do título do periódico, um PREFIXO associado a cada Coleção e apresentado no Quadro 2 adiante:

http://statbiblio.scielo.org//stat_biblio/index.php?state=15&lang=pt&country=PREFIXO&issn=ISSN&CITED%5B%5D=TITULO&YNG%5B%5D=all

A exceção a este padrão é a Coleção de Periódicos da África do Sul, que apresenta pequena diferença no *link* de acesso às estatísticas:

http://statbiblio.za.scielo.org//stat_biblio/index.php?state=15&lang=pt&country=PREFIXO&issn=ISSN&CITED%5B%5D=TITULO&YNG%5B%5D=all

Quadro 2 - Informações sobre as Coleções do SciELO: prefixos

COLEÇÃO	PREFIXO
África do Sul	sza
Argentina	arg
Brasil	scl
Chile	chl
Colômbia	col
Cuba	cub
Espanha	esp
México	mex
Portugal	org
Saúde Pública	spa
Venezuela	ven

Fonte: Mattos (2013, p.55).

2.2.3 Vinculação da(s) área(s) de conhecimento de cada periódico

O SciELO também disponibiliza um link para cada Coleção com a lista de periódicos organizada por assunto – no caso, por área de conhecimento. Para a Coleção de Saúde Pública, o link http://www.scielosp.org/scielo.php?script=sci_subject&lng=pt&nrm=iso fornece essa lista.

Outro programa foi criado para interpretar o conteúdo deste *link* e armazenar no banco de dados a(s) área(s) de conhecimento de cada periódico identificado na etapa anterior, realizando o processo diariamente para manter os dados atualizados.

De acordo com o levantamento realizado, concluiu-se que o acesso às áreas de conhecimento de cada periódico também segue o mesmo padrão identificado no Quadro 1, variando apenas uma parte do *link*, alterando-se o parâmetro *script=sci_alphabetic* para *script=sci_subject*:

http://DOMÍNIO/scielo.php?script=sci_subject&lng=pt&nrm=iso,

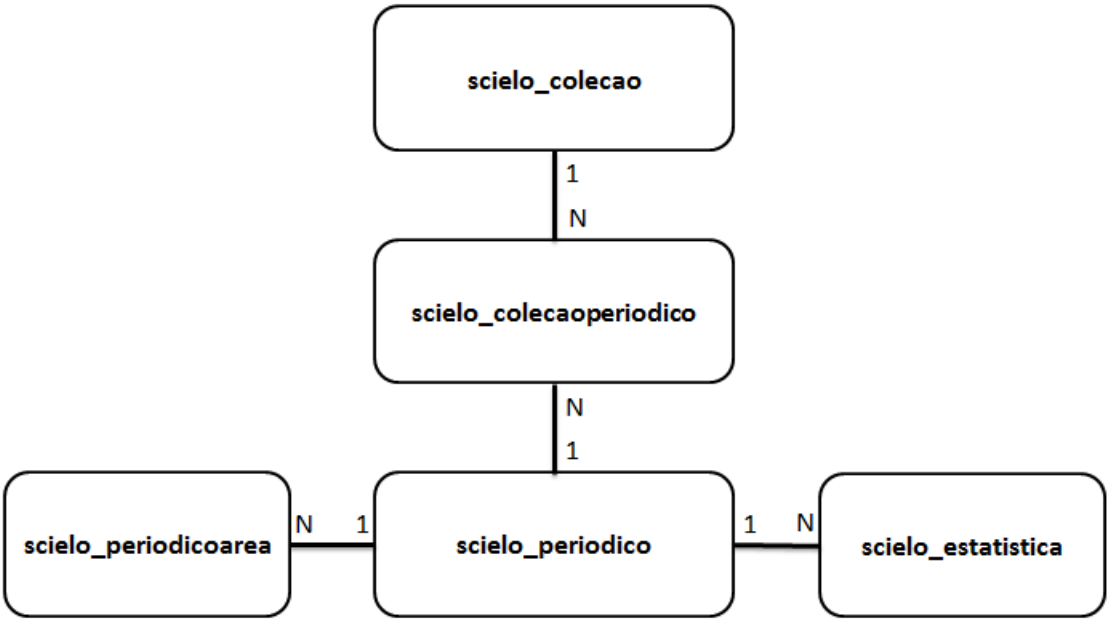
Assim, a partir da informação sobre o DOMÍNIO de cada Coleção, é possível identificar-se a sua lista de periódicos e também as respectivas áreas de

conhecimento. É importante ressaltar que o parâmetro *lng=pt* garante a padronização dos nomes das áreas em português, permitindo posteriormente o agrupamento de periódicos de diferentes Coleções por área de conhecimento.

2.2.4 Criação do Módulo "Dados Cadastrais" do banco de dados

Os dados cadastrais identificados nas etapas anteriores foram armazenados em um banco de dados, especificamente no Módulo "Dados Cadastrais". A Figura 5 apresenta o modelo relacional simplificado desse Módulo:

Figura 5 - Módulo "Dados Cadastrais"



Fonte: Mattos (2013, p. 56).

A tabela *scielo_periodico* armazena os dados de cada periódico identificado: ISSN, título, total de fascículos e situação (corrente ou não corrente). A tabela *scielo_periodicoarea* armazena a relação entre cada periódico e a(s) área(s) de conhecimento à(s) qual(is) ele está vinculado. Os dados estatísticos anuais de cada periódico (número de fascículos, total de citações concedidas e recebidas, e a média de artigos por fascículo) são armazenados na tabela *scielo_estatistica*. A tabela *scielo_colecao* armazena o nome da coleção e os *links* de acesso à lista de artigos, às estatísticas e aos arquivos XML. A relação entre cada coleção e seus periódicos é armazenada na tabela *scielo_colecaoperiodico.*

É importante ressaltar que a única tabela que tem seus dados informados manualmente é *scielo_colecao* – todas as outras são preenchidas a partir de dados coletados de forma automática e contínua.

A partir dessa análise, conclui-se que para a obtenção automática da lista de periódicos, das estatísticas e dos arquivos XML de uma Coleção do SciELO, é preciso identificar-se o DOMÍNIO e o PREFIXO de cada Coleção. Essas informações permitem a definição de um critério objetivo para a escolha das Coleções que comporão a amostra: coleções do SciELO que apresentem o *link* para acesso à lista de periódicos, às estatísticas e aos arquivos XML. Coleções sem um PREFIXO ou DOMÍNO explicitados, no caso, não serão consideradas para a composição da amostra por ser impossível a identificação automática dos *links* para acesso à lista de periódicos, estatísticas ou arquivos XML. Essas informações são apresentadas no Quadro 3 a seguir:

Quadro 3 - Informações sobre as Coleções do SciELO

COLEÇÃO	DOMÍNIO	PREFIXO	OBSERVAÇÃO
África do Sul	www.scielo.org.za	sza	
Argentina	www.scielo.org.ar	arg	
Bolívia	www.scielo.org.bo	-	SEM PREFIXO
Brasil	www.scielo.br	scl	
Brasil *Proceedings*	www.proceedings.scielo.br	-	SEM PREFIXO
Chile	www.scielo.cl	chl	
Colômbia	www.scielo.org.co	col	
Costa Rica	www.scielo.sa.cr	-	SEM PREFIXO
Cuba	scielo.sld.cu	cub	
Espanha	scielo.isciii.es	esp	
México	www.scielo.org.mx	mex	
Paraguai	scielo.iics.una.py	-	SEM PREFIXO
Peru	www.scielo.org.pe	-	SEM PREFIXO
Portugal	www.scielo.gpeari.mctes.pt	org	
Saúde Pública	www.scielosp.org	spa	
Social Sciences	socialsciences.scielo.org	-	SEM PREFIXO
Uruguai	www.scielo.edu.uy	-	SEM PREFIXO
Venezuela	www.scielo.org.ve	ven	
West Indian Medical Journal	caribbean.scielo.org	-	SEM PREFIXO

Fonte: Mattos (2013, p. 57).

A partir dessas informações definiu-se inicialmente um conjunto com 11 Coleções de Periódicos: África do Sul, Argentina, Brasil, Chile, Colômbia, Cuba, Espanha, México, Portugal, Saúde Pública e Venezuela. Posteriormente, como a

Coleção da Argentina não retornou nenhum arquivo XML, a mesma foi retirada da amostra. O resumo dos dados dessas Coleções é apresentado na Figura 6 a seguir:

Figura 6 - Coleções do SciELO utilizadas para validação da primeira fase do protótipo

	Periódicos	Fascículos		Artigos		Citações Concedidas		Citações Recebidas		Correntes	Não correntes	Arquivos processados (BC)	Citações armazenadas (BC)
11 coleções	**1090**	**31.781**	**100,00%**	**388.964**	**100,00%**	**10.338.928**	**100,00%**	**856.905**	**100,00%**	**949**	**141**	**25.003**	**515.273**
Africa do Sul	40	293	0,92%	3.101	0,80%	100.746	0,97%	1.579	0,18%	40	0	0	0
Argentina	106	1.661	5,23%	14.686	3,78%	448.716	4,34%	21.707	2,53%	104	2	255	6.672
Brasil	331	14.926	46,97%	217.423	55,90%	5.798.519	56,08%	566.869	66,15%	279	52	13.307	290.609
Chile	105	3.051	9,60%	33.463	8,60%	895.517	8,66%	44.966	5,25%	94	11	0	0
Colombia	165	2.776	8,73%	28.972	7,45%	891.242	8,62%	29.070	3,39%	165	0	756	17.241
Cuba	50	1.724	5,42%	18.840	4,84%	337.291	3,26%	27.614	3,22%	46	4	928	13.584
Espanha	56	1.912	6,02%	17.056	4,38%	479.588	4,64%	28.526	3,33%	38	18	2.514	50.240
Mexico	128	1.869	5,88%	17.148	4,41%	559.483	5,41%	25.299	2,95%	114	14	1.565	38.282
Portugal	48	776	2,44%	6.838	1,76%	91.586	0,89%	4.000	0,47%	33	15	0	0
Venezuela	56	1.304	4,10%	11.803	3,03%	298.405	2,89%	13.426	1,57%	31	25	0	0
Saude Publica	15	1.489	4,69%	19.634	5,05%	437.835	4,23%	93.849	10,95%	15	0	25.003	515.273

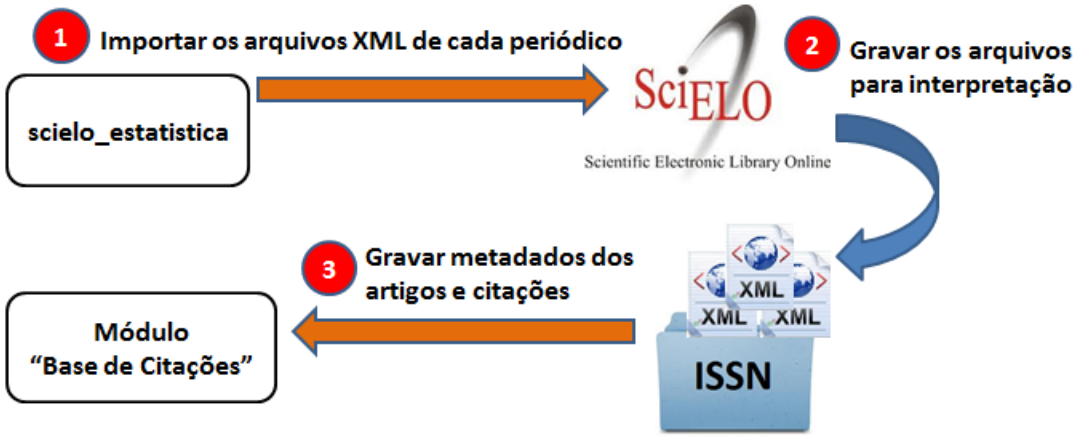
Fonte: imagem do sistema desenvolvido pelo autor[3]

[3] Disponível em: <http://cmca.srv.br/prototipo/metabuscador_index.php>. Acesso em: 10 abril 2014.

2.3 FASE II: Criação da Base de Citações

A segunda Fase proposta trata da criação da base de citações, conforme a Figura 7, e para esse processo são necessários três passos: o primeiro corresponde à identificação e armazenamento dos arquivos XML disponíveis no SciELO; o segundo, à interpretação desses arquivos para extração dos metadados e informações sobre cada referência citada; e o último, que deve armazenar todas as informações extraídas de cada arquivo XML no banco de dados – no Módulo "Base de Citações". Esses passos são descritos a seguir.

Figura 7 - FASE II: Criação da Base de Citações



Fonte: (Mattos, 2013, p. 21)

2.3.1 Levantamento dos arquivos XML de cada periódico

Para a obtenção dos arquivos XML do SciELO o primeiro passo foi a identificação do padrão de composição do link de acesso a cada arquivo da Coleção de Saúde Pública do SciELO. Esse padrão identificado diferencia-se pelo ISSN de cada periódico consultado, o ano de publicação, o número do fascículo e o sequencial do artigo:

http://www.scielosp.org/scieloOrg/php/articleXML.php?pid=IDENTIFICADOR&lang=pt

onde o IDENTIFICADOR corresponde a

S**9999-9999**9999**9999**99999
ISSN ANO NUM SEQ

Assim, S**0102-311X2013000**00019 significa:

ISSN = **0102-311X**
ANO = **2013**
NUMERO = **0007**
Sequencial dentro do número = 00019 (19º arquivo)

E o acesso para o arquivo é possível a partir do *link*:

http://www.scielosp.org/scieloOrg/php/articleXML.php?pid=*S0102-311X2013000700019*&lang=pt

A partir das informações armazenadas no Módulo "Dados Cadastrais" para cada periódico, especificamente em relação às estatísticas anuais, torna-se possível a identificação automática dos anos para os quais existem arquivos XML, quantos fascículos existem em cada ano e quantos artigos (arquivos) estão disponíveis. A partir dessas informações foi criado um programa que gera os *links* conforme o padrão acima e captura os arquivos XML – preservando como nome do arquivo o IDENTIFICADOR descrito acima – e armazena-os em um arquivo compactado e nomeado com o ISSN do periódico.

Para o acesso a múltiplas coleções, a identificação completa do *link* para o arquivo XML é dada pela expressão já descrita anteriormente, que também depende da informação sobre o DOMÍNO:

http://DOMÍNIO/scieloOrg/php/articleXML.php?pid=IDENTIFICADOR&lang=pt

2.3.2 Interpretação dos arquivos XML do SciELO

A estrutura dos arquivos XML do SciELO apresenta 2 grandes grupos de informações: dados gerais sobre o artigo, e dados específicos sobre cada referência utilizada. O QUADRO 4 apresenta as principais *tags* identificadas e o tipo de informação armazenada em cada uma delas. As *tags* listadas estavam inseridas em dois grandes grupos: um contido nas tags <front> e </front> que apresenta dados gerais do artigo, como título, periódico, volume, edição, páginas, palavras-chave e resumos; e outro contido nas tags <back> e </back> com o detalhamento de cada referência citada:

Quadro 4 - Estrutura do arquivo XML do SciELO

TAG	DESCRIÇÃO
<front>	Contém os metadados gerais do artigo
<journal-meta>	Apresenta o ISSN, título e título abreviado do periódico, e o nome do editor
<article-meta>	Contém os dados específicos do artigo: doi; título em cada idioma; nome e sobrenome dos autores; instituição dos autores; resumo em cada idioma; palavras-chave; dia, mês e ano de publicação; volume, número e páginas
<back>	Apresenta os dados de cada referência citada
<ref id="Bn">	Cada referência é agrupada dentro de uma *tag* identificada com um número n sequencial. Estão disponíveis informações sobre o tipo de citação; nome e sobrenome dos autores; título e idioma; fonte; dia, mês e ano de publicação; volume e número; páginas; editor e local.

Fonte: Mattos (2013, p. 59).

A interpretação dessas *tags* permitiu a separação dos metadados de cada arquivo e de cada citação, e a Figura 8 a seguir apresenta o resultado parcial da extração de citações de um arquivo XML do SciELO:

Figura 8 - Resultado parcial da interpretação do XML SciELO: citações de um artigo

label	citation-type	article-title-pt	article-title-en	article-title-es	collab	source	year	month	day	publisher-name	publisher-loc	conf-name	conf-loc	conf-date	page-range	volume	issue	numero
1					COMUNIDAD ANDINA	Régimen común sobre derecho de autor y derechos conexos	1993											
2						Estudio sobre las limitaciones y excepciones al derecho de autor en beneficio de bibliotecas y archivos (Comité Permanente de Derecho de Autor y Derechos Conexos, OMPI)	2008												
3	book					Towards consensus on the electronic use of publications in libraries	2001			SUB	Göttingen							
4					EIFL	Draft law on copyright. Including model exceptions and limitations for libraries and consumers. Based on WIPO draft law on copyright and related rights (version 2005)	2009											
5	journal			Derecho de autor y bibliotecas digitales: a la búsqueda del equilibrio entre intereses contrapuestos		Transinformação	2008								123-131	20	2	2
6	journal			Protección tecnológica y privilegios de las bibliotecas: regulación en la legislación de derecho de autor de los países de la Unión Europea		Nuovi Annali della Scuola Speciale per Archivisti e Bibliotecari	2009								225-240	23		
7	book			Bibliotecas y derechos de autor: análisis comparativo de la nueva legislación de España y Portugal		Información, investigación y mercado laboral en información y documentación en España y Portugal	2008			Universidad de Salamanca	Salamanca				801-811			

Fonte: Mattos (2013, p. 60)

Para o primeiro teste, os arquivos XML obtidos no passo anterior para a PCI para o ano de 2012 foram interpretados por um programa específico que, ao final do processamento, apresentou um resumo para validação com os mesmos dados estatísticos da base SciELO (Figura 9): total de fascículos, de artigos, a média de artigos por fascículo e o total de citações.

Figura 9 - Resumo da importação automática de arquivos XML: PCI 2012.

Arquivos processados: 34. Tempo total: 20.846111 segundos. Médio: 0.613122

ANO	VOLUME	EDICOES	ARTIGOS	MEDIA ARTIGOS	CITACOES
2012	17	3	34	11.3333	1038
TOTAL	-	3	34	11.3333	1038

Fonte: Mattos (2013, p. 61).

Em relação à interpretação dos arquivos XML, não foram necessários experimentos para a resolução de problemas de uso de diferentes normas para elaboração de referências, tais como a norma da Associação Brasileira de Normas Técnicas (ABNT) ou *Vancouver*. A importação de arquivos XML nesses formatos não gerou problemas de inconsistências, pois a estrutura de *tags* permaneceu inalterada. A partir dos testes realizados, infere-se que o SciELO faz o tratamento das diferenças de formatos antes da geração do arquivo XML, não havendo necessidade de tratamento desse tipo de problema.

Em relação à classificação dos materiais em cada citação, foi possível a identificação dos tipos *book*, *journal* e *confpro*, e para os demais casos essa informação não estava preenchida. Alguns exemplos são listados a seguir na Figura 10:

Figura 10 - Tipos de citações encontradas: amostra

tipocitacao	nomefonte	ano
journal	Persp. Ci. Inf.	jan.
book	Epistemologia: curso de atualização	1980
confpro	Anais...	2003
	A face oculta do documento: tradição e inovação no limiar da Ciência da	2009

Fonte: Mattos (2013, p. 6).

Conforme explicitação anterior, não é objetivo central deste trabalho a geração de dados bibliográficos, e por isso algumas sugestões ou problemas identificados – como o tipo de citação em branco e o ano "jan." na Figura 10, não foram tratados. As sugestões seguem ao final do trabalho como contribuição deste.

O primeiro exemplo, ilustrado na Figura 11 adiante, refere-se ao item "A face oculta do documento: tradição e inovação no limiar da Ciência da Informação" que é uma tese, mas não está classificada com esse tipo – uma oportunidade de melhoria na geração do arquivo XML, que foi devidamente conferido:

Figura 11 - Possível melhoria para a geração do arquivo XML: identificação de teses.

```
<ref id="B62">
<nlm-citation citation-type="">
<person-group person-group-type="author">
<name>
<surname><![CDATA[RABELLO]]></surname>
<given-names><![CDATA[R.]]></given-names>
</name>
</person-group>
<person-group person-group-type="editor">
<name>
</name>
</person-group>
<source><![CDATA[A face oculta do documento: tradição e inovação no limiar da
Ciência da Informação]]></source>
<year>2009</year>
<page-range>331p</page-range></nlm-citation>
</ref>
```

Fonte: Mattos (2013, p. 62).

O segundo exemplo, apresentado na Figura 12, sugere que os periódicos com informação de mês textual parecem não ser corretamente identificados para a geração dos arquivos XML:

Figura 12 - Possível erro na geração do arquivo XML: datas textuais

```
<ref id="B4">
<nlm-citation citation-type="journal">
<person-group person-group-type="author">
<name>
<surname><![CDATA[ARBOIT]]></surname>
<given-names><![CDATA[A. E.]]></given-names>
</name>
<name>
<surname><![CDATA[BUFREM]]></surname>
<given-names><![CDATA[L. S.]]></given-names>
</name>
<name>
<surname><![CDATA[FREITAS]]></surname>
<given-names><![CDATA[J. L.]]></given-names>
</name>
</person-group>
<article-title xml:lang="pt"><![CDATA[Configuração epistemológica da Ciência da
Informação na literatura periódica brasileira por meio de análise de citações (1972-
2008)]]></article-title>
<source><![CDATA[Persp. Ci. Inf.]]></source>
<year>jan.</year>
<month>/a</month>
<day>br</day>
<numero>1</numero>
<issue>1</issue>
<page-range>18-43</page-range><publisher-loc><![CDATA[v. 15 ]]></publisher-loc>
```

Fonte: Mattos (2013, p. 62)

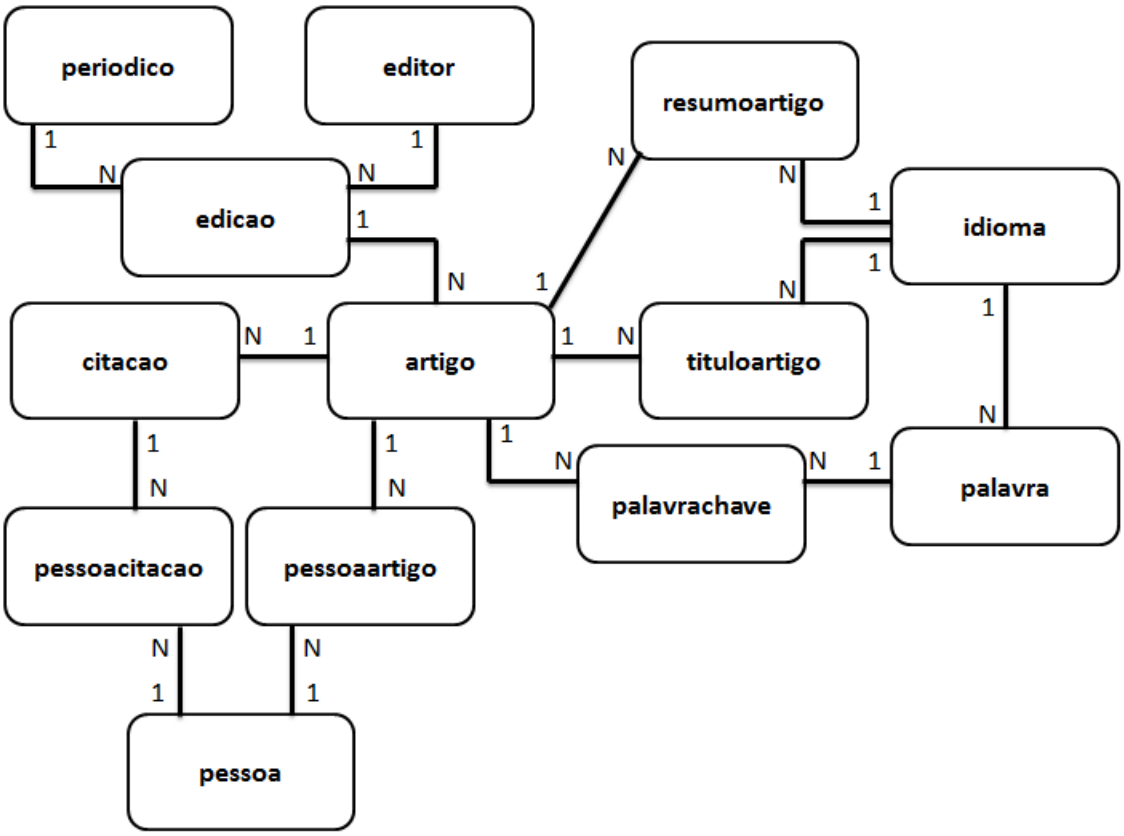
O programa que interpreta os arquivos XML armazena as informações sobre cada artigo (arquivo XML) e suas citações no banco de dados cuja estrutura é apresentada no tópico a seguir.

2.3.3 Criação do Módulo "Base de Citações" do banco de dados

Após a identificação detalhada dos dados passíveis de extração do XML, foi modelado o banco de dados para armazenar a base de citações. O desenvolvimento desse modelo foi revisado a cada novo arquivo XML importado, pois eventualmente outros campos foram identificados e o banco de dados foi adequado para o seu correto armazenamento.

Os metadados identificados na etapa anterior foram armazenados especificamente no Módulo "Base de Citações". A Figura 13 apresenta o modelo relacional simplificado desse Módulo:

Figura 13 - Módulo "Base de Citações".



Fonte: Mattos (2013, p. 63).

A tabela *periodico* armazena o título do periódico e seu ISSN; a tabela *editor* armazena o nome do editor, e a tabela *edicao* relaciona *editor* e *periodico*. Cada artigo vincula-se a uma única *edicao*, e na tabela *artigo* são armazenados o identificador e o *Digital Object Identifier* (DOI) de cada arquivo importado. O título e o resumo de cada artigo são armazenados nas respectivas tabelas (*tituloartigo* e *resumoartigo*) com a identificação do idioma (títulos em português, espanhol, inglês, francês, africanês e latim). A tabela *idioma* também identifica cada *palavra* encontrada na lista de palavras-chave dos artigos – os artigos são ligados às palavras a partir da tabela *palavrachave*.

Cada artigo armazenado possui várias citações, cujos metadados estão separados em campos específicos da tabela *citacao*. Existe uma tabela (*pessoa*) com todos os nomes e sobrenomes de pessoas identificadas, tanto nos metadados dos artigos como das citações – essas pessoas são relacionadas a cada artigo (*pessoaartigo*) ou citação (*pessoacitacao*).

3 Exemplo de Aplicação: dados da Coleção Saúde Pública

Diversos testes foram realizados com o protótipo, tanto para um único periódico como para grupos de periódicos ou coleções específicas. Até o momento, aproximadamente 400.000 arquivos XML foram capturados e interpretados. Os poucos problemas identificados que impediram a interpretação dos arquivos XML referem-se a situações que serão detalhadas em outro trabalho, e estão vinculadas a erros na estrutura XML ou no conteúdo dos dados fonte do SciELO.

Para exemplificar a riqueza de informações que pode ser disponibilizada com este protótipo, são apresentados a seguir alguns resultados da interpretação da Coleção Saúde Pública (CSP).

Em termos de quantidade de registros, a CSP apresentou os seguintes valores para cada tabela do banco de dados[4]: 14 periódicos, 14 editores, 1.335 fascículos (tabela *edicao*), 23.780 artigos (18.693 apresentaram citações associadas a eles; 5.087 não) e 491.739 citações. Foram encontrados 37.124 resumos (tabela *resumoartigo*) – 10.200 em português, 17.506 em inglês, 8.010 em espanhol e 1.408 em francês – e 44.696 títulos (tabela *tituloartigo*), dos quais 11.804 em português, 20.912 em inglês, 10.563 em espanhol, 1.416 em francês e 1 em latim.

Do total de 149.874 palavras-chave (tabela *palavrachave*) apresentadas em todos os artigos, 35.586 correspondem a termos distintos (tabela *palavra*) – sem nenhum tratamento de desambiguação – que foram apresentados em inglês (15.777), em português (9.201), em francês (1.543) e em espanhol (9.065). As 10 palavras que mais ocorreram nesta amostra foram: México (1.642), Epidemiologia e *Epidemiology* (637 cada), *Risc Factors* (615), *Mortality* (488), *Socioeconomic Factors* (475), *Public health* (474), Colombia (462), Brasil (444) e *Brazil* (438).

Foram identificados 73.859 autores de artigos (tabela *pessoaartigo*) sendo 44.595 nomes distintos (tabela *pessoa*) – sem desambiguação de nomes. Os 10 autores com mais artigos produzidos foram: Forattini, Oswaldo Paulo (128 citações), Minayo, Maria Cecília de Souza (100), Victora, Cesar G. (80), Monteiro, Carlos Augusto (79), Leal, Maria do Carmo (77), Laurenti, Ruy (73), Szwarcwald, Celia Landmann (68), Lima-Costa, Maria Fernanda (66), Barros, Marilisa Berti de Azevedo (65), Tomasi, Elaine (63).

[4] Consulta ao banco de dados realizada em 07 abril 2013; dados retirados de Mattos (2013, p. 101-2).

Dos 1.240.734 autores identificados nas citações (tabela *pessoacitacao*), 432.592 são distintos (tabela *pessoa*) – sem tratamento de desambiguação. Os 10 nomes mais citados foram: Victora, CG (1.884), Minayo, MCS (1.553), Monteiro, CA (1.278), Barros, FC (997), Szwarcwald, CL (833), Lopez, AD (697), Murray, CJL (667), Lima-Costa, MF (623), Souza, ER (599) e Leal, MC (595).

As fontes mais citadas, desconsiderados os valores "branco" (4.411 ocorrências), foram: Cad Saúde Pública (10.281), *Lancet* (7.222), Rev Saúde Pública (6.596), JAMA (3.482), BMJ (3.455), *Soc Sci Med* (2.472), *Bull World Health Organ* (2.277), *Am J Public Health* (2.210), *N Engl J Med* (2.123) e *Salud Pública Mex* (2.019).

4 Considerações Finais

O processo proposto neste trabalho busca a obtenção automática dos metadados dos artigos e referências citadas disponíveis no formato *eXtensible Markup Language* (XML) para a criação de uma base de citações considerando como fonte primária a *Scientific Electronic Library On-line* (SciELO). A automatização deste processo é um passo inicial para permitir a criação de vários índices de citações que vão desde uma base de citação geral para América Latina e Caribe (MATTOS; CENDÓN, 2013) até bases de citações descentralizadas para controle de coleções específicas como a Coleção Saúde Pública, ou de revistas específicas, como a Revista Perspectivas em Ciência da Informação ou de qualquer revista SciELO que tenha seus dados disponíveis. Estas bases de citações, uma vez implantadas e automaticamente atualizadas e disponibilizadas aos usuários finais no site das revistas permitiram a otimização da recuperação das informações dessas revistas, facilitando a resposta a questões, por exemplo, de autores e fontes mais citados.

O protótipo desenvolvido (MATTOS; CENDÓN, 2013) monitora o SciELO para identificar inclusão de periódicos e outras alterações, e captura e interpreta novos arquivos XML disponibilizados para atualização automática da base de dados de citações para um conjunto qualquer de periódicos cadastrados no SciELO.

Está em fase final a importação de todos os arquivos XML disponíveis no SciELO, além do estudo da criação de um procedimento para desambiguação dos dados gravados, vinculado à prática da graduação da Escola de Ciência da Informação (ECI) da Universidade Federal de Minas Gerais (UFMG).

Automatic generation of a citation index for SciELO using its XML files

Abstract

The paper demonstrates the use of eXtensible Markup Language (XML) files from the Scientific Electronic Library Online (SciELO) for the creation of a citation database for its journals. The focus of this article is the description of the methodology used to obtain automatic statistical data from SciELO for the journals, as well as the XML files available. These files were interpreted and metadata of articles and references used in their production were automatically recorded in a database. The Public Health Collection was used to illustrate the application of the prototype. The paper suggests the disponibilization of the database created for each journal at its own site.

Key-words: *Information Science. SciELO. Citation Index. XML.*

Referências

CENDÓN, B. V; GUIMARÃES, M. C. S.; SILVA, C. H.; OLIVEIRA, M.; MATTOS, M. C.; SANTANA, R. A. L; FERNANDES, W. R. *Construção e validação de metodologia e protótipo para criação de um índice de citações da produção científica brasileira: um estudo de caso na área de saúde coletiva*. Projeto submetido à chamada MCTI/CNPQ/MEC/CAPES nº 18/2012. 2012. (Não publicado)

CHEN, Peter. Entity-Relationship Modeling: Historical Events, Future Trends, and Lessons Learned. In: *Software Pioneers*: Contributions to Software Engineering, Broy M. and Denert, E. (eds.), Springer-Verlag, Berlin, Lecturing Notes in Computer Sciences, June 2002, p. 100-114

CHEN, Peter. The Entity-Relationship Model-Toward a Unified View of Data. *ACM Transactions on Database Systems*, Vol. 1, No. 1, March, p. 9-36. 1976

CODD, E. F. A relational model of data for large shared data banks. *Commun. ACM [S.I.]*, v. 13, n. 6, p. 377-387, 1970.

CODD, E. F. *Derivability, redundancy and consistency of relations stored in large data banks*. IBM Research Report, 1969.

CRESWELL, J. W.; CLARK, V. L. *Designing and conducting mixed methods research*. 2nd. ed. Los Angeles: Sage, 2011. 457p.

GOGOLIN, I. et al. European social science citation index: a chance for promoting European research? *European Educational Research Journal*, v.2, n.4, p.574-593, 2003.

GUIMARÃES, M. C. S; SILVA, C. H.; SANTANA, R. A. L.; BRAGA, G. M.; BOCHNER, R.; GOLDBAUM, M. *Métricas em saúde coletiva*: bases quantitativas e qualitativas para a criação de um índice de citação da literatura nacional em Saúde Coletiva. Relatório de pesquisa para o Projeto CNPq – Processo 403522/2008-0. 2011. (Não publicado)

LAKATOS, E. M.; MARCONI, M. de A. *Técnicas de pesquisa*. São Paulo: Atlas, 6ed, 2007. 289 p.

MATTOS, M. C. de. *Proposta de uma base de citações da literatura científica por meio da extração automática de dados do SciELO*. 2013. 172f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2013

MATTOS, M. C. de; CENDÓN, B. V. 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. In: 65ª Reunião Anual da SBPC – Sociedade Brasileira para o Progresso da Ciência, 2013. Recife. *Anais*... Recife: UFPE, 2013. No prelo.

MENEGHINI, R. Avaliação da produção científica e o Projeto SciELO. *Ci. Inf.*, Brasília, v. 27, n. 2, p. 219-220, maio/ago. 1998

TRIPODI, T. *et al*. *Análise da pesquisa social*. Rio de Janeiro: Francisco Alves, 1975. 337 p.

WEBSTER, B. M. Polish sociology citation index as an example of usage of national citation indexes in scientometric analysis of social sciences. *Journal of Information Science*, v. 24, n.1, p.19-32, 1998.

XIN-NING, S. et al. Developing the Chinese social science citation index. *Online Information Review*, v. 25, n. 6, p. 365-69, 2001.