XIV ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO

XIV ENANCIB 2013

GT 11 – INFORMAÇÃO & SAÚDE

A COLEÇÃO SAÚDE PÚBLICA EM NÚMEROS: 1967 A 2013

Modalidade de apresentação: Comunicação Oral

MAX CIRINO DE MATTOS

BEATRIZ VALADARES CENDÓN

A Coleção Saúde Pública em números: 1967 a 2013

Resumo: Este trabalho, produto de pesquisa em andamento, demonstra o uso de arquivos *eXtensible Markup Language* (XML) da *Scientific Electronic Library Online* (SciELO) para criação de uma base de citações das revistas da Coleção Saúde Pública (CSP), mostrando também os problemas existentes para a criação desta base e seu uso para fornecer uma visão bibliométrica da CSP, para o período em que os arquivos XML estavam disponíveis no SciELO. Inicialmente é descrita a metodologia usada para a obtenção automática dos dados estatísticos do SciELO para as revistas da CSP, bem como dos arquivos XML disponíveis. Esses arquivos foram interpretados e os metadados dos artigos e das referências usadas na sua produção foram gravados automaticamente em uma base de citações. Os resultados mostraram 18.698 artigos e 491.613 citações usadas, informação um pouco diferente daquela disponível nos dados fonte do SciELO para a CSP – 17.672 artigos e 488.252 citações. Esta diferença é explicada e soluções são sugeridas. A análise bibliométrica foi realizada sem nenhum tipo de tratamento de desambiguação dos dados. São apresentados relatórios similares aos modelos iniciais de Garfield (1972) para o *Science Citation Index* (SCI): frequências de citações, estatísticas dos periódicos citados e estatísticas dos periódicos citantes. Também são apresentados os autores mais citados, as palavras-chave mais usadas e os autores que mais produziram artigos. Sugere-se a disponibilização da base de citações, com atualização automática, de forma integrada ao site de cada uma das revistas listadas. Também é sugerida a criação de um processo de desambiguação vinculado à prática dos cursos de graduação da Escola de Ciência da Informação (ECI) da Universidade Federal de Minas Gerais (UFMG) em parceria com a Fundação Oswaldo Cruz (FIOCRUZ).

Palavras-chave: Ciência da Informação. Saúde Pública. SciELO. Base de Citações.

Title

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 the journals in the Public Health Collection (PHC), showing the problems for the creation of this database. The paper also provides a demonstration of the utility of this citation database by presenting a bibliometric view of the PHC of SciELO for the period in which the XML files were available. Initially the article describes 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. Results showed 18.698 articles and 491.613 citations, information a little bit different than that available in the data for the PHC at SciELO – 17.672 articles and 488.252 citations. This difference is interpreted and explained. Solutions are proposed. Without any disambiguation treatment for the data, reports are presented based on Garfield's initial models (1972) for the Science Citation Index (SCI): journal citation frequencies, statistics on cited journals and statistics on citing journals. Another reports show the 10 most cited authors, the most used keywords and authors who produced more articles in the journal. The paper suggests the disponibilization of the database and the creation of a disambiguation process linked to the undergraduate courses of School of Information Science at the Federal University of Minas Gerais (UFMG) in partnership with Oswaldo Cruz Foundation (FIOCRUZ).

Key-words: Information Science. Public Health. SciELO. Citation Index.

1 INTRODUÇÃO

Garfield (1972, p.527) explica que desde 1927 diversos autores – entre eles Gross e Gross, e Bradford – mapeavam partes da rede existente de periódicos científicos, mas não existia um mapa geral desses periódicos. Ele afirma que, apesar do interesse e esforço desses pesquisadores, a dificuldade prática para compilar o grande volume de dados de forma manual era o grande desafio a ser vencido. Para romper esse desafio, o autor aponta como solução o uso dos dados disponíveis em meio magnético e usados para a produção do *Science Citation Index* (SCI), que havia passado de 600 periódicos em 1964 para 2.400 em 1972 – e, no final de 1971, essa base continha mais de 27 milhões de referências. Outro trabalho do autor aponta para 15 milhões de artigos publicados desde 1945 e mais de 200 milhões de referências citadas (GARFIELD, 1992, p.2). Em 1995 ele afirma que existiam 3.300 periódicos cadastrados (GARFIELD, 1995, p.88). O site do *Institute for Scientific Information* (ISI), responsável pelo SCI, mostra que existem mais de 12.000[1] periódicos atualmente.

Inspirado pelo mesmo desafio, e buscando uma solução baseada em dados disponíveis em meio magnético, foi desenvolvido, em pesquisa em andamento, um protótipo de um sistema para a criação automática de uma base de citações brasileira de artigos da *Scientific Electronic Library Online* (SciELO). A criação automática dessa base de citações representa um passo inicial importante para a criação de uma *Web of Science* para a América Latina e Caribe (MATTOS e CENDÓN, 2013) a partir da aplicação da metodologia descrita a seguir para todos os periódicos indexados no SciELO.

O presente trabalho demonstra, utilizando como exemplo a Coleção Saúde Pública (CSP) da *Scientific Electronic Library Online* (SciELO), como os arquivos eXtensible Markup Language (XML) do SciELO podem ser utilizados para preenchimento automático do conteúdo da base de citações, mostrando também os problemas existentes para criação desta base a partir do uso destes arquivos. Apresenta também um dos possíveis usos desta base de citações fornecendo uma visão bibliométrica da CSP.

[1] Disponível em <http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/>. Acesso em 07 abr. 2013.

2 PROTÓTIPO E INTERPRETAÇÃO DOS ARQUIVOS XML DO SciELO

Para o desenvolvimento do protótipo utilizado para a criação da base de dados de citações 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 e 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*.

O protótipo desenvolvido trata da obtenção e interpretação do conteúdo dos arquivos *eXtensible Markup Language* (XML) disponíveis no SciELO. A metodologia proposta para a criação da base de citações do SciELO apresenta duas fases: (I) a obtenção de informações estatísticas anuais de cada periódico do SciELO para composição dos seus dados cadastrais, e (II) a obtenção e interpretação dos arquivos XML de cada um dos periódicos para composição da base de citações.

O SciELO apresenta um resumo estatístico para os periódicos indexados, denominado “Lista de dados fonte”, e no caso do ISSN 0036-3634 a FIG. 1 a seguir apresenta as seguintes informações:

FIGURA 1 – Lista de dados fonte na Coleção Saúde Pública: ISSN 0036-3634



Salud Pública de México
ISSN 0036-3634

Data do último processamento 11-06-2013

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 ▽
Salud pública Méx	**127**	**1279**	**43387**	**2875**	**10.07**
2013	2	18	810	69	9.00
2012	7	70	2529	219	10.00
2011	10	101	3769	246	10.10
2010	8	93	2963	276	11.62
2009	10	119	5088	333	11.90
2008	10	104	3963	285	10.40
2007	10	97	3117	244	9.70
2006	8	85	2832	227	10.62
2005	6	46	1472	145	7.67
2004	6	45	1496	124	7.50
2003	11	120	3802	186	10.91
2002	7	78	2409	78	11.14
2001	6	61	1933	67	10.17
2000	6	52	1252	68	8.67
1999	8	68	2106	123	8.50
1998	6	55	1694	101	9.17
1997	6	67	2152	84	11.17
total	**127**	**1279**	**43387**	**2875**	**10.07**

Fonte: SciELO, 2013[2]

De acordo com esse relatório, para o período definido entre 1997 e 2013, estão disponíveis 127 fascículos, 1.279 artigos e 43.387 citações para a Salud Pública de México. Esses dados foram buscados automaticamente na referida página do SciELO e gravados no banco de dados, para cada ISSN da CSP.

[2] Disponível em <http://statbiblio.scielo.org//stat_biblio/index.php?state=15&lang=pt&country=spa&issn=0036-3634&CITED[]=SALUD+PUBLICA+DE+MEXICO&YNG[]=all>. Acesso em 13 jun. 2013.

A partir dessas informações 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 para cada periódico. A partir dessas informações foi criado um programa que gera os *links* e captura os arquivos XML, armazenando-os em um arquivo compactado e nomeado com o ISSN do periódico.

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 1 a seguir 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 1 – 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: Desenvolvido pelos autores

A interpretação dessas *tags* permitiu a separação dos metadados de cada arquivo e de cada citação.

3 OBTENÇÃO DOS DADOS DO SciELO E CRIAÇÃO DA BASE DE CITAÇÕES

O protótipo foi testado na criação automática da base de citações para a Coleção Saúde Pública do SciELO, utilizando todos os arquivos XML disponíveis na data de realização dos testes.

A lista de periódicos da Coleção Saúde Pública, obtida automaticamente do SciELO, apresentou 15 periódicos:

QUADRO 2 – Periódicos da Coleção Saúde Pública do SciELO

ISSN	Título
0021-2571	Annali dell'Istituto Superiore di Sanità
0034-8910	Revista de Saúde Pública
0036-3634	Salud Pública de México
0042-9686	Bulletin of the World Health Organization
0102-311x	Cadernos de Saúde Pública
0124-0064	Revista de Salud Pública
0213-9111	Gaceta Sanitaria
0864-3466	Revista Cubana de Salud Pública
1020-4989	Revista Panamericana de Salud Pública
1135-5727	Revista Española de Salud Pública
1413-8123	Ciência e Saúde Coletiva
1415-790x	Revista Brasileira de Epidemologia
1555-7960	MEDICC Review
1726-4634	Revista Peruana de Medicina Experimental y Salud Pública
1851-8265	Salud Colectiva

Fonte: desenvolvido pelos autores

A "Lista de dados fonte" de cada um dos periódicos da Coleção Saúde Pública constantes do QUADRO 2 foi obtida automaticamente e gravada no banco de dados – a única exceção foi o periódico "*Annali dell'Istituto Superiore di Sanità*", ISSN 0021-2571, pois os dados não estavam disponíveis na data da consulta[3].

De acordo com a FIG. 2, adiante, o resumo da importação apresenta, para cada ano, o total de fascículos, de artigos, a média (de artigos em cada fascículo) e o total de citações. Essas informações são apresentadas para a base de citações (BC) e para o SciELO, tendo como fontes, respectivamente, a interpretação dos arquivos XML e a lista de dados fonte do periódico. Eventuais diferenças entre os números da BC e do SciELO são identificadas nas cores azul (quando o número da BC é superior ao do SciELO) e vermelho (se o número do SciELO for superior ao da BC). Quando há divergência entre os números da BC e do SciELO, é apresentado o percentual em relação ao número do SciELO. Caso os números da BC e do SciELO sejam idênticos, são apresentados na cor verde.

[3] Disponível em <http://statbiblio.scielo.org//stat_biblio/index.php?state=15&lang=pt&country=spa&issn=0021-2571&CITED%5B%5D=annali%20dellistituto%20superiore%20di%20sanita&YNG%5B%5D=all>. Acesso em 01 abr. 2013.

FIGURA 2 – Resumo dos dados de importação das citações da Coleção de Saúde Pública do SciELO

	FASCICULOS			ARTIGOS			MEDIA			CITAÇÕES		
ISSN	BC	SCIELO	DIFERENÇA	BC	SCIELO	DIFERENÇA	BC	SCIELO	DIFERENÇA	BC	SCIELO	DIFERENÇA
TOTAL (14)	**1.330**	**1.347**	**-17 (-1,26%)**	**18.689**	**17.672**	**1.017 (5,75%)**	**14,05**	**13,12**	**0,93 (7,09%)**	**491.613**	**488.252**	**3.361 (0,69%)**
1851-8265	26	26	0	200	165	35 (21,21%)	7,69	6,35	1,34 (21,10%)	5.820	5.833	-13 (-0,22%)
1726-4634	12	12	0	311	278	33 (11,87%)	25,92	23,17	2,75 (11,87%)	6.509	6.509	0
1555-7960	6	6	0	55	46	9 (19,57%)	9,17	7,67	1,50 (19,56%)	1.554	1.560	-6 (-0,38%)
1415-790x	55	56	-1 (-1,79%)	666	651	15 (2,30%)	12,11	11,63	0,48 (4,13%)	18.983	18.997	-14 (-0,07%)
1413-8123	84	87	-3 (-3,45%)	2.442	2.315	127 (5,49%)	29,07	26,61	2,46 (9,24%)	71.386	71.512	-126 (-0,18%)
1135-5727	96	96	0	835	805	30 (3,73%)	8,70	8,39	0,31 (3,69%)	24.944	23.697	1.247 (5,26%)
1020-4989	184	186	-2 (-1,08%)	1.674	1.769	-95 (-5,37%)	9,10	9,51	-0,41 (-4,31%)	48.705	49.858	-1.153 (-2,31%)
0864-3466	56	25	31 (124,00%)	610	334	276 (82,63%)	10,89	13,36	-2,47 (-18,49%)	11.879	7.147	4.732 (66,21%)
0213-9111	69	70	-1 (-1,43%)	1.006	851	155 (18,21%)	14,58	12,16	2,42 (19,90%)	24.235	24.534	-299 (-1,22%)
0124-0064	53	47	6 (12,77%)	628	575	53 (9,22%)	11,85	12,23	-0,38 (-3,11%)	15.391	13.949	1.442 (10,34%)
0102-311x	197	218	-21 (-9,63%)	4.001	3.955	46 (1,16%)	20,31	18,14	2,17 (11,96%)	114.408	115.388	-980 (-0,85%)
0042-9686	159	159	0	1.702	1.428	274 (19,19%)	10,70	8,98	1,72 (19,15%)	46.674	47.047	-373 (-0,79%)
0036-3634	80	98	-18 (-18,37%)	1.131	1.014	117 (11,54%)	14,14	10,35	3,79 (36,62%)	34.792	35.064	-272 (-0,78%)
0034-8910	253	261	-8 (-3,07%)	3.428	3.486	-58 (-1,66%)	13,55	13,36	0,19 (1,42%)	66.333	67.157	-824 (-1,23%)

Fonte: desenvolvida pelos autores[4]

[4] Disponível em < http://cmca.srv.br/prototipo/metabuscador_mostraresumo.php>. Acesso em 05 abr. 2013.

Três situações identificadas merecem destaque para explicar algumas das diferenças apresentadas entre os números do SciELO e os obtidos na interpretação dos arquivos XML: (I) a exclusão de informações anuais dos dados fonte do SciELO; (II) a estrutura incompleta de *tags* no arquivo XML, que resultou na não incorporação das citações; e (III) problemas na configuração das *tags* dos arquivos XML que inviabilizam o acesso aos mesmos.

As duas primeiras situações foram encontradas para o ISSN 0124-0064, da Revista de Salud Pública. É possível observar que foram interpretados 628 arquivos XML, com 15.391 citações no total:

FIGURA 3 – Estrutura das tags XML dos arquivos interpretados: ISSN 0124-0064

ESTRUTURA DOS ARTIGOS NA BC (tags XML) - 0124-0064	ARTIGOS BC	CITAÇÕES BC
article; front; /front; back; ref-list; ref-id; /ref; /ref-list; /back; /article;	498	11.902
article; front; /front; back; ref-list; ref-id; /ref; /ref-list; /back; /article; Editorial;	130	3.489
TOTAL	**628**	**15.391**

Fonte: desenvolvida pelos autores[5]

Entretanto, o resumo anual do SciELO apresentou 575 artigos e 13.949 citações, conforme a FIG. 4 a seguir. Uma explicação parcial dessa diferença é a ausência dos dados fonte para os anos de 2001 e 2002 (nesse caso, o programa apresenta todos os números do SciELO zerados). É importante ressaltar que, uma vez que existem arquivos XML interpretados para esses dois anos, deduz-se que os dados estavam disponíveis em algum momento (pois o método de importação dos arquivos XML depende dos dados fonte) e foram excluídos.

[5] Disponível em < http://cmca.srv.br/prototipo/metabuscador_mostraissn.php?issn=0124-0064>. Acesso em 05 abr. 2013.

FIGURA 4 – Resumo da importação de dados do SciELO: dados fonte X arquivos XML: ISSN 0124-0064

0124-0064	FASCICULOS			ARTIGOS			MEDIA			CITAÇÕES		
ANO	BC	SCIELO	DIFERENÇA	BC	SCIELO	DIFERENÇA	BC	SCIELO	DIFERENÇA	BC	SCIELO	DIFERENÇA
TOTAL	**53**	**47**	**6 (12,77%)**	**628**	**575**	**53 (9,22%)**	**11,85**	**12,23**	**-0,38 (-3,11%)**	**15.391**	**13.949**	**1.442 (10,34%)**
2012	3	4	-1 (-25,00%)	50	50	0	16,67	12,50	4,17 (33,36%)	1.228	1.228	0
2011	6	6	0	86	87	-1 (-1,15%)	14,33	14,50	-0,17 (-1,17%)	2.082	2.099	-17 (-0,81%)
2010	7	8	-1 (-12,50%)	101	100	1 (1,00%)	14,43	12,50	1,93 (15,44%)	2.605	2.625	-20 (-0,76%)
2009	6	6	0	89	89	0	14,83	14,83	0,00	2.208	2.208	0
2008	6	6	0	88	88	0	14,67	14,67	0,00	2.190	2.190	0
2007	4	4	0	54	54	0	13,50	13,50	0,00	1.168	1.172	-4 (-0,34%)
2006	4	5	-1 (-20,00%)	52	52	0	13,00	10,40	2,60 (25,00%)	1.246	1.003	243 (24,23%)
2005	3	3	0	28	28	0	9,33	9,33	0,00	646	646	0
2004	4	4	0	22	22	0	5,50	5,50	0,00	660	660	0
2003	3	1	2 (200,00%)	18	5	13 (260,00%)	6,00	5,00	1,00 (20,00%)	482	118	364 (308,47%)
2002	4	0	4	23	0	23	5,75	0,00	5,75	410	0	410
2001	3	0	3	17	0	17	5,67	0,00	5,67	466	0	466

Fonte: desenvolvida pelos autores[6]

[6] Disponível em < http://cmca.srv.br/prototipo/metabuscador_mostraissn.php?issn=0124-0064>. Acesso em 05 abr. 2013.

O segundo problema, a estrutura incompleta das *tags* do arquivo XML, pode ser observado na FIG. 5 a seguir:

FIGURA 5 – Falha na estrutura das tags XML dos arquivos interpretados: ISSN 0124-0064

SITUAÇÃO - 0124-0064	ESTRUTURA DOS ARTIGOS NA BC (tags XML)	ARTIGOS BC
Artigos sem citações incorporadas	article; front; /front; back; ref-list; ref-id; /ref; Referências identificadas: 17/17 Ano: 2011	1

Fonte: desenvolvida pelos autores[7]

A estrutura incompleta das *tags* (faltam */ref-list, /back e /article*) impossibilitou a importação do arquivo XML para o ano de 2011: conforme a FIG. 4, no ano de 2011 a diferença é de exatamente 1 artigo e 17 citações. A correção da estrutura de *tags*, nesse caso, permitiria a incorporação do artigo e das citações, igualando as informações apresentadas para este ano.

O terceiro problema – a indisponibilidade dos arquivos XML – ocorreu, por exemplo, com o ISSN 0036-3634, que apresentou a maior diferença: 140 artigos e 8.184 citações a menos na BC em relação ao SciELO. Essa diferença justifica-se a partir dos dados fonte listados pelo SciELO para os anos de 1997 (67 artigos, 2.152 citações), 1998 (55 artigos e 1.694 citações), 1999 (68 artigos, 2.106 citações) e 2000 (52 artigos e 1.252 citações), que totalizam 242 artigos e 7.204 citações.

Considerando o padrão identificado para os arquivos do SciELO,

o IDENTIFICADOR corresponde a S**9999-9999**9999**9999**99999
ISSN ANO NUM SEQ

Assim, S**0036-3634****1997****0001**00004 significa:

ISSN = **0036-3634**

ANO = **1997**

NUMERO = **0001**

Sequencial dentro do número = 00004 (4º arquivo)

Alguns exemplos de arquivos que não foram encontrados são listados a seguir conforme o padrão acima: S0036-36341997000100004 [8], S0036-36341997000200007 [9],

[7] Disponível em < http://cmca.srv.br/prototipo/metabuscador_mostraissn.php?issn=0124-0064>. Acesso em 01 abr. 2013.

S0036-36341998000100004 [10] , S0036-36341999000200008 [11] e S0036-36342000000200007[12]. Para todos esses arquivos, a seguinte mensagem foi apresentada:

FIGURA 6 – Erro de acesso a arquivos XML: ISSN 0036-3634



Fonte: SciELO, 2013[13]

É importante ressaltar que nova consulta realizada no dia 05/04/13 não mais identificou os dados fonte para o período de 1997 a 2000 no SciELO. Dessa forma, nova análise mostrou que as diferenças para o referido periódico foram reduzidas, como mostra a FIG. 7 a seguir:

[8] Disponível em <http://www.scielo.org.mx/scieloOrg/php/articleXML.php?pid=S0036-36341997000100004&lang=pt>. Acesso em 01 abr. 2013.

[9] Disponível em <http://www.scielo.org.mx/scieloOrg/php/articleXML.php?pid=S0036-36341997000200007&lang=pt>. Acesso em 01 abr. 2013.

[10] Disponível em <http://www.scielo.org.mx/scieloOrg/php/articleXML.php?pid=S0036-36341998000100004&lang=pt>. Acesso em 01 abr. 2013.

[11] Disponível em <http://www.scielo.org.mx/scieloOrg/php/articleXML.php?pid=S0036-36341999000200008&lang=pt>. Acesso em 01 abr. 2013.

[12] Disponível em <http://www.scielo.org.mx/scieloOrg/php/articleXML.php?pid=S0036-36342000000200007&lang=pt>. Acesso em 01 abr. 2013.

[13] Disponível em < http://www.scielo.org.mx/scieloOrg/php/articleXML.php?pid=S0036-36342000000200007&lang=pt>. Acesso em 01 abr. 2013.

FIGURA 7 – Resumo da importação de dados do SciELO: dados fonte X arquivos XML: ISSN 0036-3634

0036-3634	FASCICULOS			ARTIGOS			MEDIA			CITAÇÕES		
ANO	BC	SCIELO	DIFERENÇA	BC	SCIELO	DIFERENÇA	BC	SCIELO	DIFERENÇA	BC	SCIELO	DIFERENÇA
TOTAL	**80**	**98**	**-18 (-18,37%)**	**1.131**	**1.014**	**117 (11,54%)**	**14,14**	**10,35**	**3,79 (36,63%)**	**34.792**	**35.064**	**-272 (-0,78%)**
2012	6	6	0	78	61	17 (27,87%)	13,00	10,17	2,83 (27,83%)	2.217	2.217	0
2011	7	10	-3 (-30,00%)	133	101	32 (31,68%)	19,00	10,10	8,90 (88,12%)	3.769	3.769	0
2010	7	8	-1 (-12,50%)	105	93	12 (12,90%)	15,00	11,62	3,38 (29,09%)	2.691	2.963	-272 (-9,18%)
2009	7	10	-3 (-30,00%)	145	119	26 (21,85%)	20,71	11,90	8,81 (74,03%)	5.088	5.088	0
2008	7	10	-3 (-30,00%)	135	108	27 (25,00%)	19,29	10,80	8,49 (78,61%)	3.966	3.966	0
2007	7	10	-3 (-30,00%)	101	97	4 (4,12%)	14,43	9,70	4,73 (48,76%)	3.117	3.117	0
2006	7	8	-1 (-12,50%)	85	85	0	12,14	10,62	1,52 (14,31%)	2.832	2.832	0
2005	6	6	0	46	46	0	7,67	7,67	0,00	1.472	1.472	0
2004	6	6	0	45	45	0	7,50	7,50	0,00	1.496	1.496	0
2003	7	11	-4 (-36,36%)	120	120	0	17,14	10,91	6,23 (57,10%)	3.802	3.802	0
2002	7	7	0	78	78	0	11,14	11,14	0,00	2.409	2.409	0
2001	6	6	0	60	61	-1 (-1,64%)	10,00	10,17	-0,17 (-1,67%)	1.933	1.933	0

Fonte: desenvolvida pelos autores[14]

[14] Disponível em < http://cmca.srv.br/prototipo/metabuscador_mostraissn.php?issn=0036-3634>. Acesso em 05 abr. 2013.

4 DESCRIÇÃO DA COLEÇÃO SAÚDE PÚBLICA: A CSP EM NÚMEROS

Em termos de quantidade de registros, a Coleção Saúde Pública apresentou os seguintes valores para cada tabela do banco de dados, em relação aos periódicos citantes: 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).

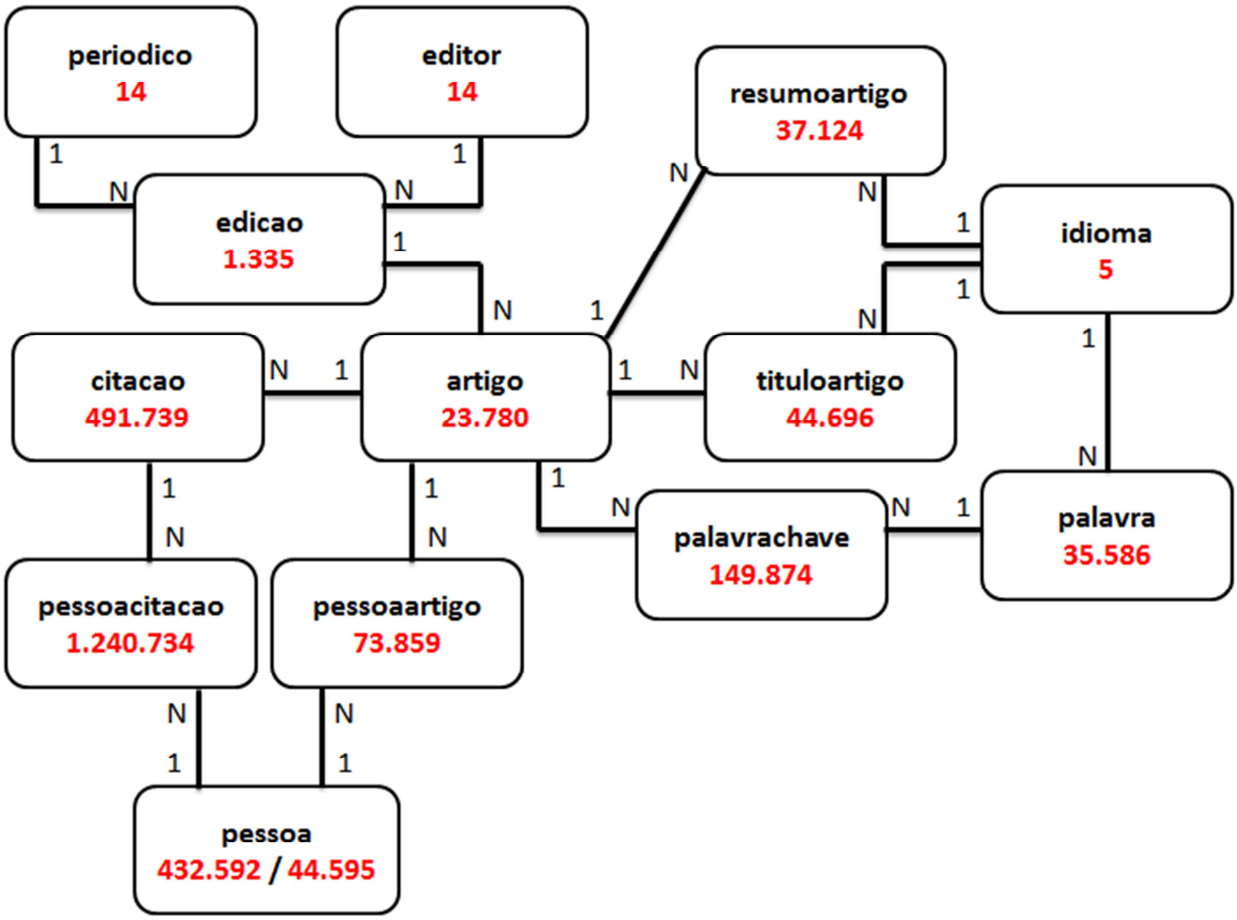
Foram identificados 73.859 autores de artigos (tabela *pessoaartigo*) sendo 44.595 nomes distintos (tabela *pessoa*) – sem desambiguação de nomes. Os 10 autores com mais artigos produzidos foram: Forattini, Oswaldo Paulo (128 citações), Minayo, Maria Cecília de Souza (100), Victora, Cesar G. (80), Monteiro, 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).

Em relação aos periódicos citados, 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, desconsiderado o valor “branco” (4.411 ocorrências) e sem nenhum tipo de recorte, 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).

A FIG. 8 a seguir apresenta os totais de registros em cada tabela do banco de dados:

FIGURA 8 – Quantidade de registros incorporados no banco de dados: Coleção Saúde Pública



Fonte: desenvolvida pelos autores[15]

As próximas FIGURAS deste tópico descrevem a amostra nos formatos dos relatórios apresentados por Garfield (1972, p.527-30): frequências de citações, estatísticas dos periódicos citados e estatísticas dos periódicos citantes. Para todos os relatórios, foram considerados os 10 periódicos mais citados listados anteriormente e detalhados os 10 últimos anos, com o total dos anos anteriores acumulados na última coluna.

O primeiro acumula o número de vezes que uma referência foi citada, e distribui essas citações por ano em que foram citadas.

O segundo, similar ao primeiro, detalha para cada fonte citada os periódicos citantes. Foi usada somente uma amostra parcial do relatório, com o periódico mais citado.

[15] Consulta ao banco de dados realizada em 07 abr. 2013.

A terceira e última lista produzida é similar à segunda, entretanto organiza os dados por periódico citante, detalhando os periódicos citados.

FIGURA 9 – Frequências de citações: Coleção Saúde Pública

FONTE CITADA	TOTAL	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	ANTERIORES
Cad Saúde Pública	10.281	280	1.486	1.729	1.281	1.207	1.326	874	788	548	404	358
Lancet	7.222	111	628	795	607	560	630	452	423	388	366	2.262
Rev Saúde Pública	6.596	151	969	932	637	602	748	557	432	422	409	737
JAMA	3.482	58	236	354	257	300	345	268	242	205	216	1.001
BMJ	3.455	36	239	332	304	365	371	270	296	245	299	698
Soc Sci Med	2.472	43	233	266	214	215	307	225	275	198	151	345
Bull World Health Organ	2.277	65	243	250	275	247	278	218	227	86	71	317
Am J Public Health	2.210	31	206	254	193	249	208	200	177	135	135	422
N Engl J Med	2.123	24	191	221	148	248	235	154	146	119	123	514
Salud Pública Mex	2.019	15	154	194	208	253	246	204	205	116	103	321

Fonte: desenvolvida pelos autores[16]

[16] Consulta ao banco de dados realizada em 10 abr. 2013.

FIGURA 10 – Estatística dos periódicos citados: Coleção Saúde Pública – visão parcial

FONTE CITADA / CITANTE	TOTAL	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	ANTERIORES
Cad Saude Publica	10.281	280	1.486	1.729	1.281	1.207	1.326	874	788	548	404	358
Cad. Saúde Pública	5.118	140	533	707	521	700	767	576	466	389	319	0
Ciênc. saúde coletiva	2.664	125	614	724	463	240	254	130	107	7	-	0
Rev. Saúde Pública	979	-	118	90	86	85	91	39	145	96	45	184
Rev. bras. epidemiol.	972	-	157	131	124	117	160	87	46	41	30	79
Rev Panam Salud Publica	341	13	51	54	49	24	36	24	7	4	6	73
Rev. salud pública	65	-	4	2	19	23	7	2	6	2	-	0
Salud pública Méx	62	-	1	11	6	14	8	8	7	-	3	4
Gac Sanit	24	-	3	8	1	1	1	3	2	3	-	2
Rev. Esp. Salud Publica	23	-	1	-	-	-	-	1	1	4	1	15
Bull World Health Organ	17	2	-	-	4	3	2	4	1	1	-	0
Rev Peru Med Exp Salud Publica	12	-	3	2	7	-	-	-	-	-	-	0
Rev Cubana Salud Pública	3	-	-	-	1	-	-	-	-	1	-	1
MEDICC rev.	1	-	1	-	-	-	-	-	-	-	-	0

Fonte: desenvolvida pelos autores[17]

[17] Consulta ao banco de dados realizada em 10 abr. 2013.

FIGURA 11 – Estatística dos periódicos citantes: Coleção Saúde Pública – visão parcial

CITANTE	FONTE CITADA	TOTAL	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	ANTERIORES
Cad. Saúde Pública		114.408	1.867	6.844	8.752	6.872	9.945	10.974	10.405	7.845	5.923	6.180	38.801
	Cad Saúde Pública	5.118	140	533	707	521	700	767	576	466	389	319	0
	Rev Saúde Pública	2.988	90	321	427	290	389	467	322	244	217	221	0
	Ciênc Saúde Coletiva	1.192	37	193	194	138	151	183	114	59	77	46	0
	Lancet	1.157	31	75	115	71	94	123	127	57	49	89	326
	Cadernos de Saúde Pública	941	-	-	-	-	-	-	-	-	-	-	941
	Revista de Saúde Pública	803	-	-	-	-	-	-	-	-	-	-	803
	Soc Sci Med	785	24	45	93	69	82	95	102	105	102	68	0
	JAMA	739	5	43	63	47	65	87	85	68	52	53	171
	BMJ	720	7	34	55	49	82	93	78	72	35	45	170
	Diário Oficial da União	708	1	46	62	63	106	138	141	61	21	26	43
	Outros	99.257	1.532	5.554	7.036	5.624	8.276	9.021	8.860	6.713	4.981	5.313	36.347

Fonte: desenvolvida pelos autores[18]

[18] Consulta ao banco de dados realizada em 10 abr. 2013.

De forma similar aos relatórios apresentados anteriormente, as FIGURAS a seguir apresentam os autores mais citados, as palavras-chave mais usadas e os autores que mais produziram artigos no periódico. Foram considerados os 10 autores mais citados, classificados em ordem alfabética.

FIGURA 12 – Autores mais citados: Coleção Saúde Pública

AUTOR CITADO	TOTAL	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	ANTERIORES
Victora, CG	1.988	57	163	215	215	146	379	136	157	126	110	284
Minayo, MCS	1.634	87	228	180	193	211	167	104	160	104	60	140
Monteiro, CA	1.329	44	152	180	154	150	111	112	56	58	83	229
Barros, FC	1.065	26	84	105	98	54	232	68	89	64	75	170
Szwarcwald, CL	866	41	99	132	81	74	94	103	94	56	49	43
Lopez, AD	710	8	63	62	84	76	65	47	48	47	37	173
Murray, CJL	681	2	32	47	51	72	34	70	51	48	53	221
Souza, ER	660	65	86	64	64	100	57	27	99	28	13	57
Lima-Costa, MF	644	17	107	166	62	71	65	51	31	40	28	6
Leal, MC	607	21	80	71	64	53	64	63	53	49	53	36

Fonte: desenvolvida pelos autores[19]

[19] Consulta ao banco de dados realizada em 10 abr. 2013.

FIGURA 13 – Palavras-chave mais utilizadas: Coleção Saúde Pública

PALAVRA CITADA	TOTAL	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	ANTERIORES
México (Espanhol)	818	2	67	59	77	94	79	88	70	42	42	198
Mexico (Inglês)	817	2	65	61	76	96	76	89	68	42	45	197
Epidemiology (Inglês)	654	14	43	54	58	43	40	26	35	22	23	296
Risk factors (Inglês)	629	9	55	42	34	48	62	45	38	52	51	193
Mortality (Inglês)	493	13	35	40	30	41	29	20	34	35	29	187
socioeconomic factors (Inglês)	492	1	34	33	34	49	38	27	23	32	36	185
Public health (Inglês)	478	5	44	52	48	39	37	21	32	19	22	159
Brazil (Inglês)	470	6	47	58	47	63	62	57	32	14	15	69
Epidemiologia (Português)	447	9	27	31	28	19	23	15	14	11	15	255
Fatores de risco (Português)	429	7	37	34	27	37	50	35	29	25	26	122

Fonte: desenvolvida pelos autores[20]

[20] Consulta ao banco de dados realizada em 10 abr. 2013.

FIGURA 14 – Autores que mais produziram: Coleção Saúde Pública (visão parcial)

CITANTE	AUTOR PRODUTOR	TOTAL	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	ANTERIORES
Cad. Saúde Pública (0102-311X)		14.822	810	1.023	1.122	1.050	1.257	1.434	1.288	963	762	712	4.401
	Lima-Costa, Maria Fernanda	49	2	2	17	2	4	1	5	3	4	4	5
	Szwarcwald, Celia Landmann	48	1	1	5	1	2	2	1	-	11	4	20
	Victora, Cesar G.	45	-	-	-	6	1	13	-	-	-	2	23
	Leal, Maria do Carmo	44	1	3	2	-	-	2	1	5	2	12	16
	Barros, Fernando C.	37	1	-	1	4	-	11	1	-	1	-	18
	Barros, Marilisa Berti de Azevedo	37	5	8	9	2	3	2	2	3	-	-	3
	Carvalho, Marilia Sá	36	2	-	4	1	1	2	4	4	4	5	9
	Olinto, Maria Teresa Anselmo	34	1	2	3	6	2	4	5	5	2	-	4
	Hallal, Pedro C.	33	-	-	2	20	5	2	2	-	2	-	0
	Santos, Iná S.	33	3	-	2	2	3	14	3	-	3	1	2
	Outros	14.426	794	1.007	1.077	1.006	1.236	1.381	1.264	943	733	684	4.301

Fonte: desenvolvida pelos autores[21]

[21] Consulta ao banco de dados realizada em 10 abr. 2013.

5 CONSIDERAÇÕES FINAIS

O protótipo desenvolvido (MATTOS e 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, não só da CSP, aqui exemplificada, como também dos outros periódicos do SciELO.

Para a realização de estudos mais detalhados, sugere-se a disponibilização da base de citações, permitindo o acesso às consultas por autor, palavra-chave e outros metadados disponíveis na base criada.

Está em estudo a criação de um procedimento para desambiguação dos dados vinculado à prática da graduação da Escola de Ciência da Informação (ECI) da Universidade Federal de Minas Gerais (UFMG), em parceria com a Fundação Oswaldo Cruz (FIOCRUZ).

REFERÊNCIAS

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)

MATTOS, Max Cirino de; CENDÓN, Beatriz Valadares. 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.

GARFIELD, E. Citation analysis as a tool in journal evaluation. *Science*, v. 178, p. 471-79, 1972.

GARFIELD, E.; WELLJAMS-DOROF A. Citation data: their use as quantitative indicators for science and technology evaluation and policy-making. *Science and Public Policy*, v. 19, p. 321-7, 1992.

GARFIELD, E. Quantitative analysis of the scientific literature and its implications for science policymaking in Latin America and the Caribbean. *Bull Pan Am Health Organ*, v. 29, p. 87-95, 1995.