Análise automática de citações disponíveis em arquivos XML da SciELO: o periódico "Perspectivas em Ciência da Informação" em números

Max Cirino de Mattos

Doutor em Ciência da Informação, Escola de Ciência da Informação, UFMG .Bolsista de pós-doutorado na ECI-UFMG

Beatriz Valadares Cendón

Ph.D em Ciência da Informação pela University of Texas at Austin, EUA. Professora Titular, Escola de Ciência da Informação, UFMG

O artigo demonstra o uso de arquivos eXtensible Markup Language (XML) da Scientific Electronic Library Online (SciELO) para a criação de uma base de citações do periódico Perspectivas em Ciência da Informação. Demonstra também o uso desta base de citações fornecendo uma visão bibliométrica desse periódico para o período em que os arquivos estavam disponíveis na SciELO: 2006 a 2014, do volume 11 edição 1 ao volume 19 edição 4, incluindo o número especial desse volume. Foram analisados 532 artigos, e os resultados mostraram 387 artigos com 10.266 citações usadas (145 artigos não apresentaram nenhuma citação). O principal objetivo do artigo é destacar a possibilidade de automação da análise bibliométrica a partir da metodologia proposta – e por esse motivo os dados são apresentados sem nenhum tipo de tratamento de desambiguação. 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. Sugere-se a disponibilização da base de citações, com atualização automática, de forma integrada ao site do periódico, e a aplicação da metodologia para outros periódicos indexados na SciELO.

Palavras-chave*: Ciência da Informação; SciELO; Base de citações; Análise de citações.*

The journal "Perspectives on Information Science" in numbers: 2006-2014

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 journal "Perspectivas em Ciência da Informação". A demonstration of the utility of this citation database is presented by providing a bibliometric view of the journal for the period in which the XML files were available in SciELO (from 2006 to 2014). 532 papers were analyzed and results showed 387 papers using 10.266 citations (145 papers without citations associated). The aim of this article is to highlight the possibility of automation of bibliometric analysis from the proposed methodology - and therefore the data are presented without any disambiguation treatment. 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 cited journals. The application of the methodology to other journals indexed in SciELO is suggested.

Keywords: *Information Science; SciELO; Citation Index; Citation analysis.*

Recebido em 10.07.2014Aceito em 21.01.2015

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 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 existirem 3.300 periódicos cadastrados

(GARFIELD, 1995, p.88). O site do ISI, responsável pelo SCI, mostra 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 automação desse processo representa um passo inicial importante para a criação de uma base de citações para a América Latina e Caribe (MATTOS; CENDÓN, 2013) a partir da aplicação da metodologia descrita a seguir para todos os periódicos indexados na SciELO.

O presente artigo demonstra como os arquivos *eXtensible Markup Language* (XML) da SciELO podem ser utilizados para preenchimento automático do conteúdo da base de citações para o periódico Perspectivas em Ciência da Informação (PCI), representante do estrato Qualis A1 da área de Sociais Aplicadas I da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e uma das poucas revistas da área de Ciência da Informação (CI) indexada na *Scientific Electronic Library Online* (SciELO) à época. Foram usados os artigos do periódico PCI – publicados entre 2006 e 2014, do volume 11 edição 1 até o volume 19 edição 4, incluindo a edição especial – disponíveis na data da coleta de dados. Demonstra também um dos possíveis usos desta base de citações fornecendo uma visão bibliométrica da PCI.

É importante ressaltar que não se pretende avaliar ou corrigir o conteúdo dos arquivos XML importados, nem realizar análises qualitativas das informações apresentadas – apenas demonstrar a possibilidade de criação de uma base de citações atualizada automática e continuamente para a PCI.

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 *et al.*, 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

[1] Disponível em: <http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/>. Acesso em: 7 abr. 2013.

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 na SciELO. A metodologia proposta para a criação da base de citações da SciELO apresenta duas fases: (I) a obtenção de informações estatísticas anuais de cada periódico da 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.

A SciELO apresenta um resumo estatístico para os periódicos indexados, denominado "Lista de dados fonte", e no caso da PCI a Figura 1 a seguir apresenta as seguintes informações:

PERSPECTIVAS

Last processed date
Jan-05-2015

Perspectivas em Ciência da Informação
ISSN 1413-9936

Source data list

∇ - click to select the sort column
Δ - indicate current order

journal title/year Δ	no. of issues ∇	no. of articles ∇	no. of granted citations ∇	no. of received citations ∇	average articles per issue ∇
Perspect. ciênc. inf	**31**	**373**	**9995**	**410**	**12.03**
2014	3	36	1073	51	12.00
2013	4	46	1121	55	11.50
2012	4	45	1414	64	11.25
2011	4	50	1365	54	12.50
2010	3	39	1192	58	13.00
2009	4	55	1353	39	13.75
2008	3	39	1050	39	13.00
2007	3	35	873	24	11.67
2006	3	28	554	26	9.33
total	**31**	**373**	**9995**	**410**	**12.03**

Figura 1 – Imagem da lista de dados fonte: Perspectivas em Ciência da Informação

Fonte: SciELO, 2015[2].

[2]Disponível em:<http://statbiblio.scielo.org//stat_biblio/index.php?state=15&lang=pt&country=scl&issn=1413-9936&CITED[]=1413-9936&YNG[]=all>. Acesso em: 27 jan. 2015.

De acordo com esse relatório, para o período definido entre 2006 e 2014, estão disponíveis 31 fascículos, 373 artigos e 9.995 citações para a PCI. Esses dados são buscados automaticamente na referida página da SciELO e gravados no banco de dados.

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. Em seguida foi criado outro 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 da SciELO apresenta dois grandes grupos de informações: dados gerais sobre o artigo, e dados específicos sobre cada referência utilizada. O Quadro 1 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 da 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: Dados da pesquisa.

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

Os dados obtidos da SciELO, de acordo com a Figura 1, foram usados para comparar os resultados obtidos no protótipo, que são detalhados na Tabela 1:

Tabela 1 – Resultados da importação XML: PCI 2006 a 2014

ANO	EDIÇÕES	ARTIGOS	MÉDIA DE ARTIGOS	CITAÇÕES
2014	5	48	9,60	1.073
2013	4	47	11,75	1.121
2012	4	45	11,25	1.414
2011	4	50	12,50	1.365
2010	3	40	13,33	1.192
2009	4	55	13,75	1.353
2008	3	39	13,00	1.050
2007	3	35	11,67	873
2006	3	28	9,33	554
TOTAL	**33**	**387**	**11,73**	**10.266**

Fonte: Dados da pesquisa.

O protótipo apresentou 33 fascículos (edições), 387 artigos e 10.266 citações, considerando apenas artigos que possuem ao menos uma citação. Uma das diferenças observadas ocorreu no total de artigos para o ano de 2010, com 1 artigo a menos no relatório do SciELO. Os arquivos XML foram conferidos para este ano, e nova pesquisa manual ao SciELO mostrou que uma resenha[3] apresenta exatamente sete referências – entretanto, esta resenha não está incluída no total de 39 artigos do ano de 2010. Outra diferença refere-se a alguns fascículos de 2014 que possuem arquivos XML disponíveis, mas os dados estatísticos parecem não estar atualizados.

Cada um dos 387 arquivos XML da PCI foi interpretado e as informações foram armazenadas no Módulo "Base de Citações". A comparação anual entre os dados obtidos automaticamente da SciELO e as informações armazenadas a partir da interpretação dos arquivos XML é apresentada a seguir.

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 a 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 da SciELO são identificadas nas cores azul (quando o número da BC é superior ao da SciELO) e vermelho (se o número da SciELO for superior ao da BC). Quando há divergência entre os números da BC e da SciELO, é apresentado o percentual em relação ao número da SciELO. Caso os

[3] GIL LEIVA, I. Manual de indización: teoría y práctica. Gijón: Ediciones Trea, 2008. Resenha de: BOCCATO, V. R. C. Perspect. ciênc. inf., Belo Horizonte, v. 15, n. 3, 2010. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362010000300014&lng=pt&nrm=iso>. Acesso em: 1 jan. 2013.

números da BC e da SciELO sejam idênticos, são apresentados na cor verde.

O *link* existente no número de artigos de determinado ano também permite a visualização de cada um dos artigos.

A seguir são apresentadas as Figuras 2, 3 e 4, correspondentes respectivamente ao resumo da importação de dados da SciELO, a alguns artigos do ano de 2014 para o periódico PCI, e às referências de um artigo de 2013:

1413-9936	FASCICULOS			ARTIGOS			MEDIA			CITAÇÕES		
ANO	BC	SCIELO	DIFERENÇA	BC	SCIELO	DIFERENÇA	BC	SCIELO	DIFERENÇA	BC	SCIELO	DIFERENÇA
TOTAL	**33**	**31**	**2 (6,45%)**	**387**	**373**	**14 (3,75%)**	**11,73**	**12,03**	**-0,30 (-2,49%)**	**10.266**	**9.995**	**271 (2,71%)**
2014.	5	3	2 (66,67%)	48	36	12 (33,33%)	9,60	12,00	-2,40 (-20,00%)	1.344	1.073	271 (25,26%)
2013.	4	4	0	47	46	1 (2,17%)	11,75	11,50	0,25 (2,17%)	1.121	1.121	0
2012.	4	4	0	45	45	0	11,25	11,25	0,00	1.414	1.414	0
2011.	4	4	0	50	50	0	12,50	12,50	0,00	1.365	1.365	0
2010.	3	3	0	40	39	1 (2,56%)	13,33	13,00	0,33 (2,54%)	1.192	1.192	0
2009.	4	4	0	55	55	0	13,75	13,75	0,00	1.353	1.353	0
2008.	3	3	0	39	39	0	13,00	13,00	0,00	1.050	1.050	0
2007.	3	3	0	35	35	0	11,67	11,67	0,00	873	873	0
2006.	3	3	0	28	28	0	9,33	9,33	0,00	554	554	0

Figura 2 – Imagem gerada automaticamente – Resumo da importação de dados da SciELO: dados fonte e arquivos XML

Fonte: Dados da pesquisa[4].

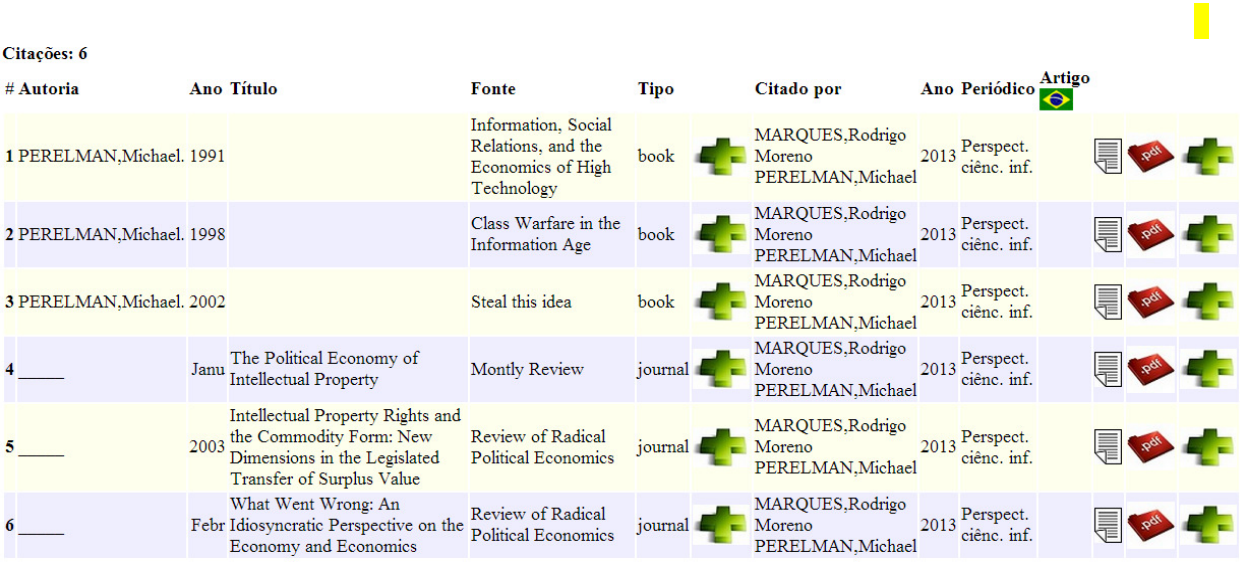
Artigos: 12

#	Autoria	Ano	Periódico	Titulo	Volume	Número	Páginas	Referênci
1	FERNANDEZ-RAMOS, Andres	2014	Perspect. ciênc. inf.		19	1	115-129	40
2	JABBOUR, Charbel Jose Chiappetta FIORINI, Paula de Camargo	2014	Perspect. ciênc. inf.	Análise do apoio dos sistemas de informação para as práticas de gestão ambiental em empresas com ISO 14001 - estudo de múltiplos casos	19	1	51-74	40
3	OLIVEIRA, Nivaldo CASTRO, Cleber Carvalho de SOUZA, Donizeti Leandro de	2014	Perspect. ciênc. inf.	Análise sociométrica da rede de relacionamento das bibliotecas que constituem o Consórcio das Universidades Federais do Sul-Sudeste de Minas Gerais	19	1	130-148	31
4	DAMASCENO, Andreia Cristina MESQUITA, Jose Marcos Carvalho de	2014	Perspect. ciênc. inf.	Atributos determinantes da baixa utilização de bibliotecas: estudo em uma instituição de ensino pública federal	19	1	149-169	32
5	SILVA JUNIOR, Jobson Francisco da AQUINO, Mirian de Albuquerque SILVA, Leyde Klebia Rodrigues da	2014	Perspect. ciênc. inf.	Comunidades virtuais de música como subsídio para a construção da identidade afrodescendente	19	1	75-89	31
6	MALUCELLI, Andreia TACLA, Cesar Augusto ZAHRA, Faruk Mustafa ROBERTO, Ademir	2014	Perspect. ciênc. inf.	Ferramentas para aprendizagem de ontologias a partir de textos	19	1	03-21	30
7	BORTOLIN, Sueli ALMEIDA JUNIOR, Oswaldo Francisco	2014	Perspect. ciênc. inf.	Mediação da literatura para leitores-ouvintes	19	1	207-226	21
8	GARCIA TAMARGO, Marco Antonio FOMBONA CADAVIECO, Javier GOULAO, Maria de Fatima	2014	Perspect. ciênc. inf.	Melhorar a atratividade da informação através do uso da realidade aumentada	19	1	37-50	43
9	SANTOS, Placida Leopoldina Ventura Amorim da Costa JORENTE, Maria Jose Vicentini	2014	Perspect. ciênc. inf.	Mídias de informação e comunicação e Ciência da Informação	19	1	190-206	20
10	DROESCHER, Fernanda Dias SILVA, Edna Lucia da	2014	Perspect. ciênc. inf.	O pesquisador e a produção científica	19	1	170-189	32
11	NASCIMENTO, Thiago Cavalcante AZEVEDO, Alexandra Katarina de VIEIRA, Leonor Laurentina ARAUJO, Richard Medeiros	2014	Perspect. ciênc. inf.	Periódicos em ação: um estudo exploratório-bibliométrico na área de Administração, Ciências Contábeis e Turismo	19	1	90-114	37
12	CUNHA, Francisco Jose Aragao Pedroza RIBEIRO, Nubia Moura PEREIRA, Hernane Borges de Barros	2014	Perspect. ciênc. inf.	Técnicas de gerenciamento de informações em uma rede de hospitais	19	1	22-36	22

Figura 3 – Imagem gerada automaticamente – Artigos listados do ISSN 1413-9936 para o ano de 2014

Fonte: Dados da pesquisa[5].

[4] Disponível em: <http://cmca.srv.br/prototipo/metabuscador_mostraissn.php?issn=1413-9936>. Acesso em: 27 jan. 2015.



Citações: 6

#	Autoria	Ano	Título	Fonte	Tipo	Citado por	Ano	Periódico	Artigo
1	PERELMAN,Michael.	1991		Information, Social Relations, and the Economics of High Technology	book	MARQUES,Rodrigo Moreno PERELMAN,Michael	2013	Perspect. ciênc. inf.	
2	PERELMAN,Michael.	1998		Class Warfare in the Information Age	book	MARQUES,Rodrigo Moreno PERELMAN,Michael	2013	Perspect. ciênc. inf.	
3	PERELMAN,Michael.	2002		Steal this idea	book	MARQUES,Rodrigo Moreno PERELMAN,Michael	2013	Perspect. ciênc. inf.	
4	_____	Janu	The Political Economy of Intellectual Property	Montly Review	journal	MARQUES,Rodrigo Moreno PERELMAN,Michael	2013	Perspect. ciênc. inf.	
5	_____	2003	Intellectual Property Rights and the Commodity Form: New Dimensions in the Legislated Transfer of Surplus Value	Review of Radical Political Economics	journal	MARQUES,Rodrigo Moreno PERELMAN,Michael	2013	Perspect. ciênc. inf.	
6	_____	Febr	What Went Wrong: An Idiosyncratic Perspective on the Economy and Economics	Review of Radical Political Economics	journal	MARQUES,Rodrigo Moreno PERELMAN,Michael	2013	Perspect. ciênc. inf.	

Figura 4 – Imagem gerada automaticamente – Referências citadas em um artigo de 2013 do ISSN 1413-9936
Fonte: Dados da pesquisa [6].

Para auxiliar na validação das informações, o programa desenvolvido também apresenta um resumo das estruturas de *tags* identificadas para o periódico:

ESTRUTURA DOS ARTIGOS NA BC (tags XML) - 1413-9936	ARTIGOS BC	CITAÇÕES BC
article; front; /front; back; ref-list; ref-id; /ref; /ref-list; /back; /article;	387	10.266
TOTAL	**387**	**10.266**

Figura 5 – Imagem gerada automaticamente – Estrutura das tags XML dos arquivos interpretados
Fonte: Dados da pesquisa[7].

A Figura 5 permite identificar 387 artigos incorporados à base de citações que possuem, no total, 10.266 citações associadas – informação idêntica à da Figura 2.

Além desses dados, também é apresentado um resumo dos arquivos incorporados à BC, porém sem citações associadas:

SITUAÇÃO - 1413-9936	ESTRUTURA DOS ARTIGOS NA BC (tags XML)	ARTIGOS BC	CITAÇÕES BC
Artigos sem citações incorporadas	article; front; /front; /article; Referências identificadas: 0/0 Ano: 2014	5	0
Artigos sem citações incorporadas	article; front; /front; /article; Referências identificadas: 0/0 Ano: 2013	16	0
Artigos sem citações incorporadas	article; front; /front; /article; Referências identificadas: 0/0 Ano: 2012	6	0
Artigos sem citações incorporadas	article; front; /front; /article; Referências identificadas: 0/0 Ano: 2011	15	0
Artigos sem citações incorporadas	article; front; /front; /article; Referências identificadas: 0/0 Ano: 2010	23	0
Artigos sem citações incorporadas	article; front; /front; /article; Referências identificadas: 0/0 Ano: 2009	23	0
Artigos sem citações incorporadas	article; front; /front; /article; Referências identificadas: 0/0 Ano: 2008	26	0
Artigos sem citações incorporadas	article; front; /front; /article; Referências identificadas: 0/0 Ano: 2007	22	0
Artigos sem citações incorporadas	article; front; /front; /article; Referências identificadas: 0/0 Ano: 2006	9	0

Figura 6 – Imagem gerada automaticamente – Arquivos XML incorporados sem citações associadas

Fonte: Dados da pesquisa[8].

[5] Disponível em: <http://cmca.srv.br/prototipo/metabuscador_mostraissn.php?issn=1413-9936>. Acesso em: 10 jul. 2014.
[6] Disponível em: <http://cmca.srv.br/prototipo/metabuscador_mostraissn.php?issn=1413-9936>. Acesso em: 10 jul. 2014.
[7] Disponível em: <http://cmca.srv.br/prototipo/metabuscador_mostraissn.php?issn=1413-9936>. Acesso em: 27 jan. 2015.
[8] Disponível em: <http://cmca.srv.br/prototipo/metabuscador_mostraissn.php?issn=1413-9936>. Acesso em: 27 jan. 2014.

Para o caso da PCI, foram encontrados 145 artigos nessa situação. Percebe-se, pela estrutura de tags apresentada, que não existe a identificação das referências nestes artigos às tags back, ref-list, ref-id, /back, /ref-list e /ref-id. Finalizando a validação da importação, cada um desses 135 artigos pode ser acessado para confirmação, sendo possível inclusive o acesso ao resumo e ao texto completo na base SciELO. A Figura 7 apresenta os seis artigos nessa situação para o ano de 2012:

Artigos: 6

#	Autoria	Ano	Periódico	Título	Volume	Número	Páginas	Referências
1		2012	Perspect. ciênc. inf.		17	1	273-274	0
2	NEVES, Jorge Tadeu de Ramos MELO, Marlene Oliveira Teixeira de	2012	Perspect. ciênc. inf.		17	3	01-01	0
3		2012	Perspect. ciênc. inf.	Editorial	17	1	01-01	0
4	NEVES, Jorge Tadeu MELO, Marlene Oliveira Teixeira de	2012	Perspect. ciênc. inf.	Editorial	17	2	1-1	0
5	MELO, Jorge Tadeu Neves e Marlene Oliveira Teixeira de	2012	Perspect. ciênc. inf.	Editorial	17	4	1-1	0
6	CARVALHO, Maria da Conceicao	2012	Perspect. ciênc. inf.	Livros & Telas. MARTINS, Aracy Alves et al. (Orgs.). Belo Horizonte: Ed. UFMG, 2011. 261 p	17	1	265-266	0

Figura 7 – Imagem gerada automaticamente – Relação de artigos sem citações associadas, na PCI, para o ano de 2012

Fonte: Dados da pesquisa[9].

4 Descrição da amostra: a PCI em números

Em termos de quantidade de registros, a PCI apresentou os seguintes valores para cada tabela do banco de dados, em relação aos periódicos citantes: 1 periódico (PCI), 1 editor (Escola de Ciência da Informação da UFMG), 33 fascículos (tabela *edicao*), 532 artigos e 10.266 citações. Foram encontrados 774 resumos (tabela *resumoartigo*) – 376 em português, 378 em inglês, 15 em espanhol e cinco em francês – e 898 títulos (tabela *tituloartigo*), dos quais 500 em português, 376 em inglês, 17 em espanhol e cinco em francês. A Figura 8 a seguir apresenta os totais de registros em cada tabela:

[9] Disponível em: <http://www.cmca.srv.br/prototipo/metabuscador_mostraartigo.php>. Acesso em: 10 jul. 2014.

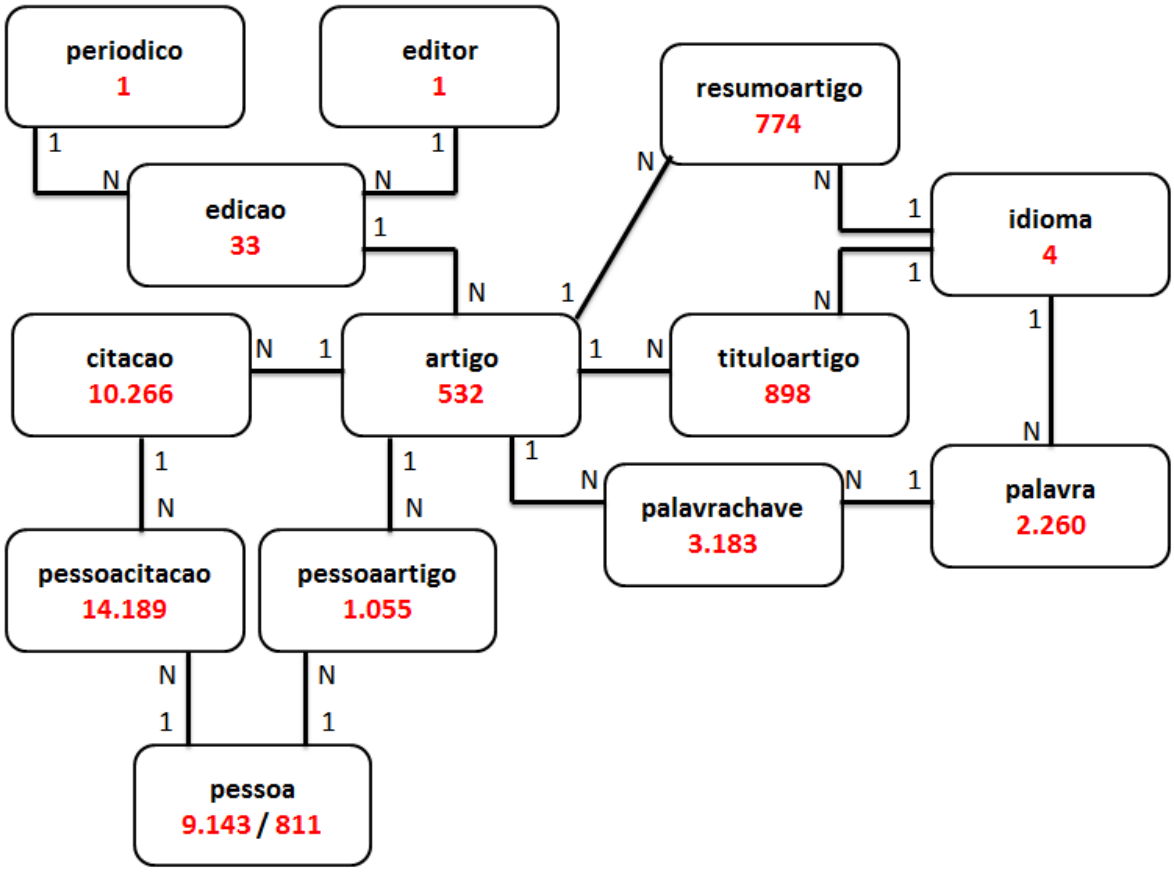


Figura 8 – Quantidade de registros incorporados no banco de dados: PCI

Fonte: Dados da pesquisa[10].

Do total de 3.183 palavras-chave (tabela *palavrachave*) apresentadas em todos os artigos, 2.260 correspondem a termos distintos (tabela *palavra*) que foram apresentados em inglês (1.102), em português (1.061), em francês (28) e em espanhol (69). As 10 palavras que mais ocorreram nesta amostra – independente do idioma – foram: Ciência da Informação (32), Information Science (29), Gestão do Conhecimento e *Knowledge Management* (23 cada), Bibliometria (21), Informação, *Information* e Internet (17 cada), Produção Científica (15) e Gestão da Informação (14).

Foram identificados 1.055 autores de artigos (tabela *pessoaartigo*) sendo 811 nomes distintos (tabela *pessoa*). Os 10 autores com mais artigos produzidos[11] – ordenados pela quantidade decrescente de artigos e ordem alfabética do sobrenome – foram: Bufrem, Leilah Santiago e Fujita , Mariângela Spotti Lopes (seis artigos cada); Cunha, Murilo Bastos da, e Todesco , José Leomar (cinco artigos cada); Araujo , Carlos Alberto Ávila; Boccato , Vera Regina Casari; Cafe, Ligia; Cendón, Beatriz Valadares; De Sordi, José Osvaldo; Dumont, Ligia Maria Moreira (cada um com quatro artigos).

[10] Consulta ao banco de dados realizada em 27 jan. 2015.

[11] Somente considerados os artigos que possuem ao menos uma citação associada; critério adotado para desconsiderar os editoriais.

Em relação aos periódicos citados, dos 14.189 autores identificados nas citações (tabela *pessoacitacao*), 9.143 são distintos (tabela *pessoa*). Os 10 nomes mais citados foram: Nonaka, I. (84 ocorrências); Castells, M. e Pinheiro, L. V. R. (39 cada); Choo, C. W. (38); Capurro, R. e Saracevic, T. (35 cada); Takeuchi, H. (34); Fujita, M. S. L. e Levy, P. (33 cada); Davenport, T. H. (32).

As fontes mais citadas, desconsiderados os valores "branco" (140 ocorrências), "Anais..." (262), "Anais" (85) e "*Proceedings*..." (56), foram: Ciência da Informação (330), Perspectivas em Ciência da Informação (158), *Scientometrics* (103), Transinformação (64), *Journal of the American Society for Information Science* (61), *Journal of Documentation* (59) e DataGramaZero (58).

A Figura 9 a seguir descreve a amostra no formato dos relatórios apresentados por Garfield (1972, p.527-30): frequências de citações, estatísticas dos periódicos citados e estatísticas dos periódicos citantes. Conforme o autor, o primeiro relatório 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; a terceira e última lista é similar à segunda, entretanto organiza os dados por periódico citante, detalhando os periódicos citados. No caso específico de um único periódico avaliado – PCI – as três listas são muito semelhantes, e por isso é apresentada a frequência de citações. 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.

CITANTE / FONTE CITADA	TOTAL	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	ANTERIORES
Perspect. ciênc. inf. (1413-9936)	10.266	-	1.344	1.121	1.414	1.365	1.192	1.353	1.050	873	554	0
Ciência da Informação	330	-	27	19	55	21	38	52	45	40	33	0
Anais...	262	-	37	27	29	41	31	31	15	7	44	0
Perspectivas em Ciência da Informação	158	-	25	19	18	21	30	13	19	7	6	0
	140	-	14	10	33	18	9	13	25	11	7	0
Scientometrics	103	-	18	13	16	6	13	-	23	9	5	0
Anais	85	-	24	4	1	3	9	30	-	9	5	0
Transinformação	64	-	3	6	13	11	12	5	3	4	7	0
Journal of the American Society for Information Science	61	-	4	10	6	6	3	3	6	16	7	0
Journal of Documentation	59	-	10	1	6	8	12	5	8	7	2	0
DataGramaZero	58	-	1	10	21	8	3	7	4	1	3	0
Outras (5.893)	8.946	-	1.181	1.002	1.216	1.222	1.032	1.194	902	762	435	0

Figura 9 – Imagem gerada automaticamente – Frequências de citações: PCI

Fonte: Dados da pesquisa[12].

De forma similar ao relatório apresentado anteriormente, as FIGURAS 10, 11, 12 e 13 a seguir apresentam os autores mais citados, as palavras-chave mais usadas, os autores que mais produziram artigos no

[12] Consulta ao banco de dados realizada em 27 jan. 2015.

periódico e as instituições dos autores produtores, sendo considerados apenas os artigos que possuem ao menos uma citação associada. Foram consideradas as 10 ocorrências mais citadas, classificadas em ordem alfabética.

CITANTE | AUTOR CITADO	TOTAL	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	ANTERIORES
Perspect. ciênc. inf. (1413-9936)	14.189	-	2.153	1.521	1.754	1.966	1.694	1.784	1.487	1.094	736	0
NONAKA, I.	84	-	12	4	3	25	10	8	13	7	2	0
CASTELLS, M.	39	-	3	4	3	8	5	3	6	6	1	0
PINHEIRO, L. V. R.	39	-	2	5	6	10	7	1	4	2	2	0
CHOO, C. W.	38	-	1	3	3	5	6	6	9	5	-	0
CAPURRO, R.	35	-	2	4	5	5	4	-	5	10	-	0
SARACEVIC, T.	35	-	1	7	3	5	7	1	7	1	3	0
TAKEUCHI, H.	34	-	2	4	1	3	8	3	6	4	3	0
FUJITA, M. S. L.	33	-	3	3	15	1	5	-	-	1	5	0
LEVY, P.	33	-	1	7	3	6	5	3	4	2	2	0
DAVENPORT, T. H.	32	-	3	1	4	3	4	7	2	6	2	0
Outros (9.133)	13.787	-	2.123	1.479	1.708	1.895	1.633	1.752	1.431	1.050	716	0

Figura 10 – Imagem gerada automaticamente – Autores mais citados: PCI

Fonte: Dados da pesquisa[13].

CITANTE | PALAVRA-CHAVE	TOTAL	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	ANTERIORES
Perspect. ciênc. inf. (1413-9936)	3.183	-	409	402	379	396	335	452	332	259	219	0
ciencia da informacao (Português)	32	-	2	3	5	5	4	4	3	2	4	0
information science (Inglês)	29	-	2	3	5	5	2	4	2	2	4	0
gestao do conhecimento (Português)	23	-	3	2	1	1	4	4	2	4	2	0
knowledge management (Inglês)	23	-	3	2	1	2	4	3	2	4	2	0
bibliometria (Português)	20	-	1	5	3	1	2	1	2	4	1	0
informacao (Português)	17	-	1	2	-	3	-	4	3	1	3	0
information (Inglês)	17	-	1	2	-	3	-	4	3	1	3	0
producao cientifica (Português)	15	-	2	1	2	2	1	2	2	2	1	0
gestao da informacao (Português)	14	-	2	2	-	1	2	1	2	4	-	0
information management (Inglês)	14	-	2	2	-	1	2	1	2	4	-	0
Outras (2.250)	2.979	-	390	378	362	372	314	424	309	231	199	0

Figura 11 – Imagem gerada automaticamente – Palavras-chave mais utilizadas: PCI

Fonte: Dados da pesquisa[14,15].

[13] Consulta ao banco de dados realizada em: 27 jan. 2015.
[14] Consulta ao banco de dados realizada em: 27 jan. 2015.
[15] A palavra "Internet", que foi uma das mais citadas com 17 ocorrências, não apareceu nessa listagem porque a mesma separa os termos por idioma: ela foi citada uma vez em espanhol, oito em inglês e oito em português.

CITANTE / AUTOR PRODUTOR	TOTAL	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	ANTERIORES
Perspect. ciênc. inf. (1413-9936)	895	-	106	116	97	124	89	114	94	79	76	0
Bufrem, Leilah Santiago	6	-	-	-	-	2	2	1	-	1	-	0
Fujita, Mariangela Spotti Lopes	6	-	1	-	2	-	1	-	-	-	2	0
Cunha, Murilo Bastos da	5	-	1	-	1	1	-	-	1	1	-	0
Todesco, Jose Leomar	5	-	1	1	-	1	2	-	-	-	-	0
Araujo, Carlos Alberto Avila	4	-	-	-	-	1	1	-	1	1	-	0
Boccato, Vera Regina Casari	4	-	-	-	-	-	2	1	-	-	1	0
Cafe, Ligia	4	-	1	-	-	1	-	1	1	-	-	0
Cendon, Beatriz Valadares	4	-	-	1	-	1	1	-	1	-	-	0
De Sordi, Jose Osvaldo	4	-	-	1	-	1	-	1	1	-	-	0
Dumont, Ligia Maria Moreira	4	-	1	-	-	-	-	2	1	-	-	0
Outros (713)	849	-	101	113	94	116	80	108	88	76	73	0

Figura 12 – Imagem gerada automaticamente – Autores que mais produziram: PCI

Fonte: Dados da pesquisa[16].

CITANTE / INSTITUIÇÃO PRODUTORA	TOTAL	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	ANTERIORES
Perspect. ciênc. inf. (1413-9936)	962	-	130	101	85	122	87	134	99	97	107	0
UFMG Escola de Ciência da Informação	29	-	2	1	1	6	4	3	9	3	-	0
Universidade Federal de Santa Catarina	18	-	8	1	2	1	1	3	1	-	1	0
Universidade de Brasilia	17	-	2	3	2	5	-	1	1	3	-	0
UFMG	16	-	-	2	1	2	-	5	3	1	2	0
Universidade Federal da Paraíba	10	-	3	-	-	2	-	-	5	-	-	0
UFMG ECI	9	-	1	-	-	-	2	1	-	1	4	0
Universidade Federal do Paraná	9	-	1	-	3	1	2	-	-	1	1	0
USP Sistema Integrado de Bibliotecas	9	-	-	-	-	-	-	-	-	1	8	0
Universidade Federal de Santa Catarina Departamento de Ciência da Informação	8	-	4	-	1	2	-	1	-	-	-	0
Universidade Federal de Santa Catarina Programa de Pós-Graduação em Ciência da Informação	8	-	2	-	1	2	-	3	-	-	-	0
Outras (555)	829	-	107	94	74	101	78	117	80	87	91	0

Figura 13 – Imagem gerada automaticamente – Instituições que mais produziram: PCI

Fonte: Dados da pesquisa[17].

Em relação à Figura 13, é importante ressaltar a necessidade de desambiguação dos dados importados dos arquivos XML. Apesar de o processo automatizado oferecer uma série de informações instantâneas, os dados obtidos não são padronizados – ou seja, a informação das instituições dos autores é totalmente livre e subjetiva.

Assim, percebe-se a ocorrência, entre as 10 instituições mais citadas da Figura 13, de "UFMG Escola de Ciência da Informação", "UFMG" e "UFMG ECI" – todos os registros representantes da UFMG.

[16] Consulta ao banco de dados realizada em: 27 jan. 2015.
[17] Consulta ao banco de dados realizada em: 27 jan. 2015.

Apenas para ilustrar as divergências que podem ocorrer em relação aos dados obtidos diretamente dos arquivos XML, foi realizada uma desambiguação manual simples das instituições, desconsiderando departamentos ou nomes por extenso – que foram substituídos pela sigla das instituições. A nova distribuição é apresentada na Figura 14 adiante:

CITANTE INSTITUIÇÃO PRODUTORA	TOTAL	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	ANTERIORES
Perspect. ciênc. inf. (1413-9936)	962	-	130	101	85	122	87	134	99	97	107	0
UFMG	132	-	5	14	7	12	11	29	18	19	17	0
UFSC	55	-	15	7	8	6	2	9	2	1	5	0
UnB	32	-	3	5	4	7	-	2	2	5	4	0
UFPB	27	-	7	-	1	4	2	-	6	6	1	0
USP	18	-	2	1	-	3	1	2	-	6	3	0
UEL	17	-	2	1	2	-	-	5	2	4	1	0
UFPR	14	-	4	1	3	1	2	1	-	1	1	0
UNESP	13	-	2	2	1	3	-	-	2	1	2	0
Fiocruz	10	-	-	-	-	-	-	1	2	7	-	0
UFRN	7	-	1	4	-	-	-	-	-	-	2	0
Outras (461)	637	-	89	66	59	86	69	85	65	47	71	0

Figura 14 – Imagem gerada automaticamente – Instituições que mais produziram: PCI – Resultado após desambiguação

Fonte: Dados da pesquisa[18].

5 Considerações finais

O protótipo desenvolvido (MATTOS; CENDÓN, 2013) monitora a SciELO para identificar inclusão de periódicos e outras alterações, captura e interpreta novos arquivos XML disponibilizados para atualização automática da base de dados de citações, não só da revista Perspectivas em Ciência da Informação, aqui exemplificada, como também dos outros periódicos do SciELO (MATTOS; CENDÓN, 2014).

No resultado geral para a PCI, entre os autores mais citados e as palavras-chave mais usadas nos artigos analisados, destacou-se o tema Gestão do Conhecimento. Para a realização de estudos mais detalhados, sugere-se a disponibilização da base de citações, com atualização automática contínua, de forma integrada ao site da PCI, permitindo o acesso às consultas por autor, palavra-chave e outros metadados disponíveis na base criada.

A metodologia retratada no artigo pode contribuir para uma análise bibliométrica automatizada dos diversos periódicos indexados na base SciELO, agilizando o processo de obtenção dos dados e oferecendo aos estudiosos do tema informações em meio digital que pode ser exportada facilmente para excel ou outras ferramentas.

Cabe ressaltar que, apesar de não se deter nesse tópico, os autores estão cientes de que os dados extraídos diretamente dos arquivos XML devem ser revisados e em alguma medida desambiguados, como ficou demonstrado na apresentação das instituições de origem dos autores dos artigos da PCI.

[18] Consulta ao banco de dados realizada em: 27 jan. 2015.

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).

Referências

GARFIELD, E. Citation analysis as a tool in journal evaluation. *Science*, n. 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*, n. 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*, n. 29, p. 87-95, 1995.

GUIMARÃES, M. C. S. *et al. 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, 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:* REUNIÃO ANUAL DA SBPC – SOCIEDADE BRASILEIRA PARA O PROGRESSO DA CIÊNCIA, 65., 2013. Recife. *Anais*... Recife: UFPE, 2013.

MATTOS, M. C. de; CENDÓN, B. V. Criação automática de uma base de citações para o SciELO a partir dos seus arquivos XML. *Informação & Tecnologia*, Marília/João Pessoa, v. 1, n. 1, p.42-67, 2014.