8ºEBBC
Encontro Brasileiro de Bibliometria e Cientometria
MACEIÓ2022
20, 21 E 22 · JULHO

ÍNDICE DA COBERTURA TEMÁTICA DE REVISTAS CIENTÍFICAS:
estudo de caso na área de Ciência da Informação

Ronnie Fagundes de Brito[1]

Resumo: Periódicos possibilitam a publicação de resultados de pesquisa, atuando sobre temáticas específicas, sendo a abrangência em sua área um elemento relevante no estudo de sua produção. Mas como qualificar a produção de um periódico em relação a sua abrangência em determinada temática? Nesse sentido, foram estruturadas métricas sobre cobertura temática numa proposta de índice de cobertura temática baseada no uso de tesauro e tratamento lexicográfico de conteúdos. A proposta é aplicada em revistas da Ciência da Informação, calculando a aderência de sua produção ao tesauro utilizado na análise e verificando associações entre as revistas em função dos temas nelas abordados.

Palavras-Chave: Frequência de palavras. Produtividade de periódicos. Indicadores de C&T.

1 INTRODUÇÃO SOBRE UM ÍNDICE DE COBERTURA TEMÁTICA COMO MÉTRICA DE PRODUÇÃO

A ciência é dinâmica e constantemente aborda novas técnicas, métodos e aspectos da realidade visando resolver problemas e melhorar a qualidade de vida. O surgimento e desaparecimento de tópicos e temas constitui um aspecto da evolução da pesquisa científica. Nesse contexto surgem questões sobre como qualificar e quantificar a abrangência ou cobertura temática de uma revista de disseminação científica.

A cobertura temática constitui um aspecto da política editorial das revistas, declarando a seu público quais as áreas/temas/tópicos em que deseja publicar ou disponibilizar conteúdos. Como exemplo, a base indexadora Scielo Brasil apresenta como critério de admissão de periódicos a apresentação de sua cobertura temática de acordo com a classificação da CAPES.

Visando descrever a cobertura temática de um revista, Curiel, Lorenzo e Hernández Acosta (2009) analisam a distribuição temática dos artigos da revista "Avanzada Científica" entrevistando membros de seu comitê editorial para definição da futura cobertura temática de sua revista. A cobertura temática também pode ser inferida pela linguagem documentária utilizada, refletindo os conceitos utilizados em determinado contexto, como no caso do

[1] Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)

repositório da Escola Nacional de Administração Pública (ENAP), onde permite observar os limites de atuação da escola. Nessa linguagem são utilizados terminologias como o Tesauro de Administração Pública, Vocabulário Controlado de Governo Eletrônico e a Classificação Bibliográfica de sua biblioteca (MULLER; OLIVEIRA, 2015).

Como métrica, a cobertura temática pode ser analisada como um índice acumulativo que demonstra a diversificação dos artigos de determinado periódico, como aponta (ALVAREZ-OSSORIO, 1997) ao descrever a cobertura temática de uma revista de documentação científica diante uma classificação de temas compartilhada pela área. Em sua análise os autores demonstram a evolução na quantidade de artigos publicados em cada tema em dois períodos distintos.

Ao ser considerada como métrica, de Moya-Anegón *et al.* (2007) analisam as bases Scopus e Ulrich's Core em relação a sua cobertura temática, geográfica e em relação a seus editores e idiomas. Nesse estudo traçam um perfil de cada base em relação aos temas tratados em seus documentos, destacando-se a multidisciplinaridade da base Scopus. Em outro estudo bibliométrico, Zacca-Gonzalez *et al.* (2014) analisaram periódicos da América Latina da área da saúde utilizando o Thematic Specialization Index (TSI). Este índice deriva do Activity Index, o qual considera a relação entre o total global de publicações em determinada área e o total de publicações que determinado país realizou nesta mesma área.

A cobertura de temas ou tópicos é também área de pesquisa em sistemas de recuperação de informação, especificamente em modelos de aprendizagem não supervisionada, os quais geram listas quantificadas de termos e documentos de modo a descrevê-los de forma automatizada (KORENČIĆ *et al.*, 2021). Por sua vez, Coronini e Magematin (1999) descrevem a análise da cobertura temática das atividades de um departamento de ensino de Ciências Sociais em uma universidade por meio de tratamento lexicográfico de sua produção técnica e científica, onde descrevem a coerência entre suas atividades de ensino e pesquisa, notando necessidade de maior conexão entre estas duas.

Diante desse contexto, onde um índice de cobertura temática serviria para acompanhar a política editorial de periódicos, busca-se avaliar o alinhamento dos temas abordados em revistas da área de Ciência da Informação ao tesauro da área. Assim como analisar o agrupamento das revistas em função das temáticas que elas abordam.

2 MÉTODOS

O acompanhamento do comportamento da produção por meio da análise de cobertura temática ou de tópicos permite analisar a aderência da publicação às temáticas relevantes à área de pesquisa.

Para a análise foram utilizados os textos em portugues de artigos publicados pelas revistas registradas na Base de Dados em Ciência da Informação (BRAPCI) no ano de 2021. Os metadados desses artigos foram coletados por meio de protocolo OAI-PMH e os textos completos foram obtidos por meio de raspagem de dados com posterior conversão para formato texto. Foi adotado o período de 2017 a 2021 para a amostra dos artigos.

No estudo foi utilizado um script Jupyter Notebook (BRITO, 2022) de modo a possibilitar a execução iterativa e experimental dos algoritmos de manipulação de dados, assim como o processamento de linguagem natural, que busca elaborar meios para a representação de textos de modo a permitir sua análise e processamento computacional numa ampla gama de aplicações (LIDDY, 2001)

Como parâmetro de referência para a análise de cobertura temática foi usado o Tesauro Brasileiro de Ciência da Informação (PINHEIRO; FERREZ, 2014), o qual permitiu filtrar os termos relacionados com a área. Para sua utilização, este foi convertido em grafo a partir do seu registro no sistema de gerenciamento de vocabulários Tematres. O grafo foi constituído por elos entre termos e categorias, termos gerais e termos específicos, e serve como base para a visualização e cálculo da cobertura temática em cada revista.

De modo a normalizar os termos extraídos dos textos é realizado o tratamento lexicográfico do corpus, removendo-se palavras irrelevantes (stop words) com seguinte lematização. Visando-se o alinhamento com os termos extraídos dos artigos, da mesma forma, os termos do tesauro também passaram por normalização/lematização.

Para a extração dos termos mais relevantes de cada artigo, de modo a ponderar termos repetidos e distingui-los dos mais significativos mas em menor quantidade foi aplicada a técnica de ponderação TF-IDF (Term Frequency - Inverse Document Frequency) como em (VUOTTO; FERNANDEZ; BOGETTI, 2015) e (AIZAWA, 2003).

Em seguida é populado um grafo com os dados de cada revista a partir do grafo do tesauro, atribuindo-se o tamanho de seus nós/nodos em função da frequência de termos extraídos de seus artigos. Nessa etapa, os nodos referentes a termos do tesauro não encontrados no corpus das revistas foram suprimidos.

De modo a quantificar a similaridade entre o grafo do tesauro e o da cobertura das revistas é calculada a distância entre estes, e visando mensurar a similaridade entre as coberturas temáticas das revistas, a distância entre os grafos de cada revista também é calculada. Para o cálculo da distância foi utilizado a "Distância de Edição do Gráfo" (GAO *et al.*, 2010).

Finalmente o grafo de cada revista é apresentado visualmente, o qual é formado pelos termos do tesauro com a ponderação da frequência de termos nos corpora analisados, assim como são plotados um dendograma de classificação das revistas em função da sua cobertura temática e também um mapa de calor com os agrupamentos das revistas em função das distâncias entre si. O quadro 1 representa o fluxo de procedimentos utilizado no estudo.

Quadro 1 - Etapas do algoritmo para análise da cobertura temática dos periódicos

1. Converte tesauro PDF para texto estruturado;
2. Cria grafo de termos do tesauros;
3. Para cada revista:
    1. Coleta OAI-PMH;
    2. Extração de metadados;
    3. Download e conversão dos pdfs dos artigos para texto;
    4. Extração de termos dos artigos;
    5. Criação do grafo com nodos para cada termo do tesauro e tamanho em função da quantidade de termos extraídos dos artigos;
    6. Geração de visualização;
    7. Calculo da distancia entre o grafo da revista com o grafo do tesauro;
    8. Calculo da distancia entre o grafo da revista e outras revistas da base;
4. Apresentação e análise de resultados.

Fonte: Elaborado pelo autor.

Cabe ressaltar que na análise não foram usadas as palavras-chave do campo dc:subject pois muitos artigos não apresentavam esse dado de forma estruturada e a extração automatizada tem se mostrado tão efetiva quanto palavras-chaves informadas pelo autor no contexto de análises bibliométricas (ZHANG *et al.*, 2016).

3 RESULTADOS

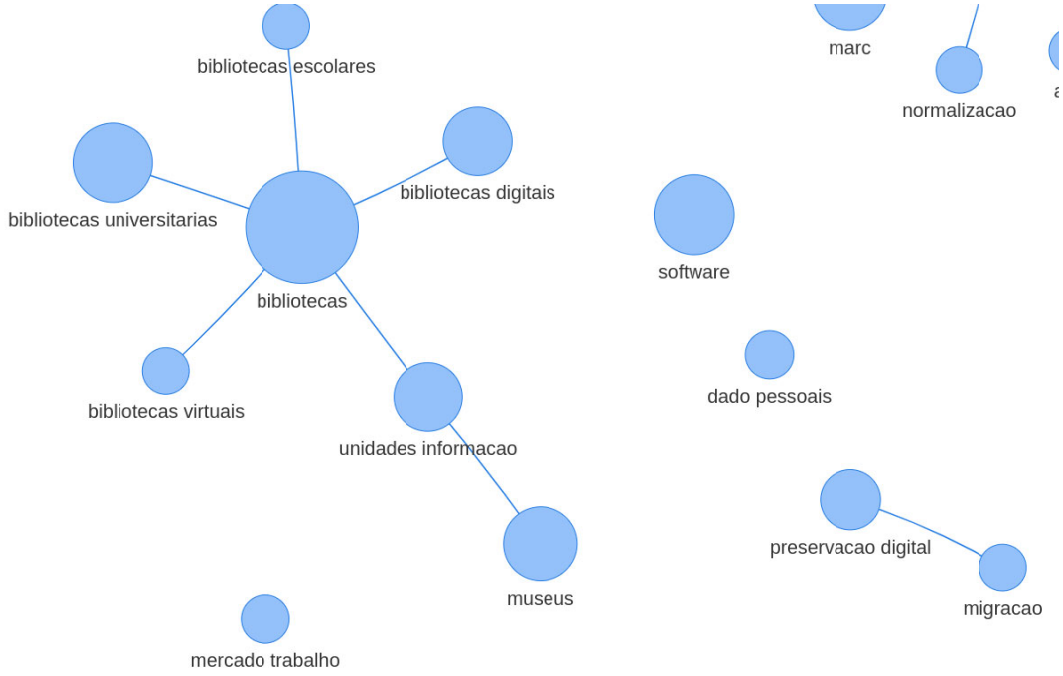
Das 52 revistas que estavam na BRAPCI em 2021, 33 puderam ser coletadas por meio da interface OAI-PMH. Dessas, a Ciência da Informação não pode ser completamente coletada devido a problemas na codificação dos metadados.

Diante a faixa temporal adotada, entre 2017 e 2021, foi possível extrair o texto completo de um total de 9111 documentos. Na extração das palavras-chave, foram extraídos termos compostos por até três palavras, sendo adotados os 15 termos mais relevantes. A "Revista

Em Questão" apresentou a maior quantidade de artigos, com 412 documentos processados, seguida da Ciência da Informação, com 369 documentos analisados. Dados de outros revistas podem ser consultados no quadro do Anexo 1.

Foi adotado um recorte adicional selecionando as revistas com mais de 30 nodos no grafo, visto que grafos pequenos tendem a não ser representativos e gerariam distorções na análise. Um recorte do grafo de cobertura temática da revista Ciência da Informação pode ser visto na figura 1. Uma versão completa dos grafos de cada revista pode ser visualizada em Brito (2022a).

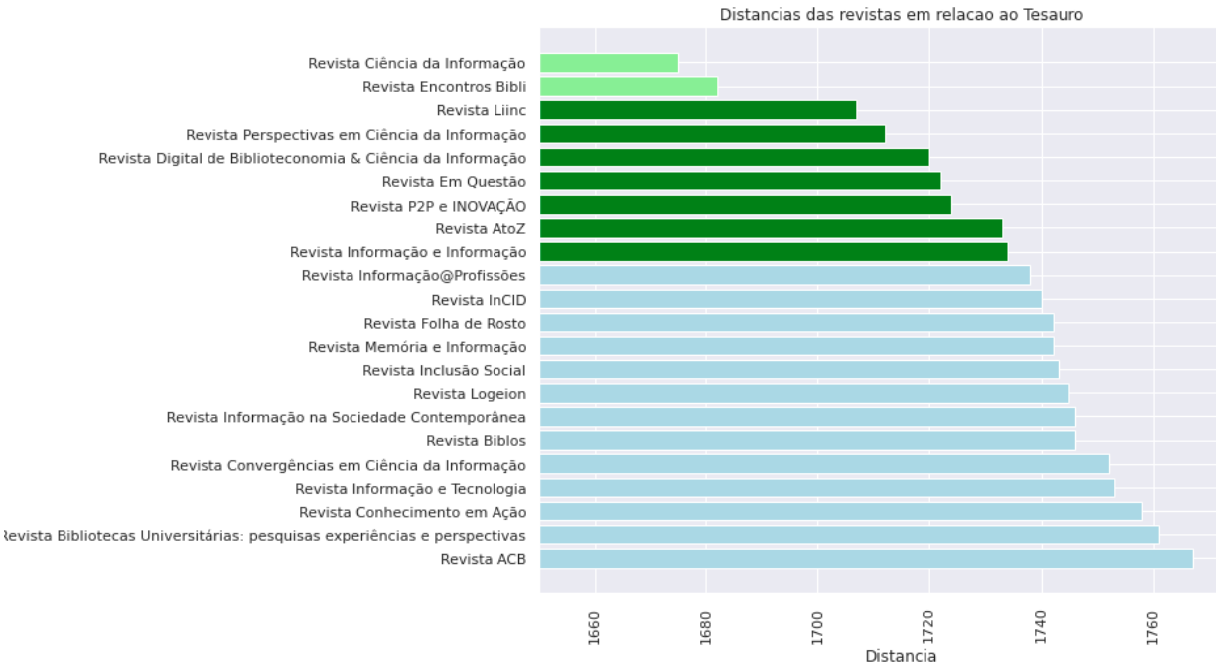
Figura 1 - Recorte de grafo representando a cobertura temática das revista Ciência da Informação



Fonte: Elaborado pelo autor.

A visualização em si possibilita reflexões a respeito dos temas mais abordados pelas revistas, contudo a informação que permite comparar as revistas entre si é gerada por meio da métrica baseada na Distância de Edição do Grafo. No cálculo dessa distância se considerou apenas a existência de associação dos documentos aos termos do tesouro, não sendo considerada a frequência de termos. As distâncias calculadas entre os grafos de cada revista e o grafo do tesauro são apresentadas na figura 2.

Figura 2 - Distância entre os grafos das revistas e o grafo do tesauro

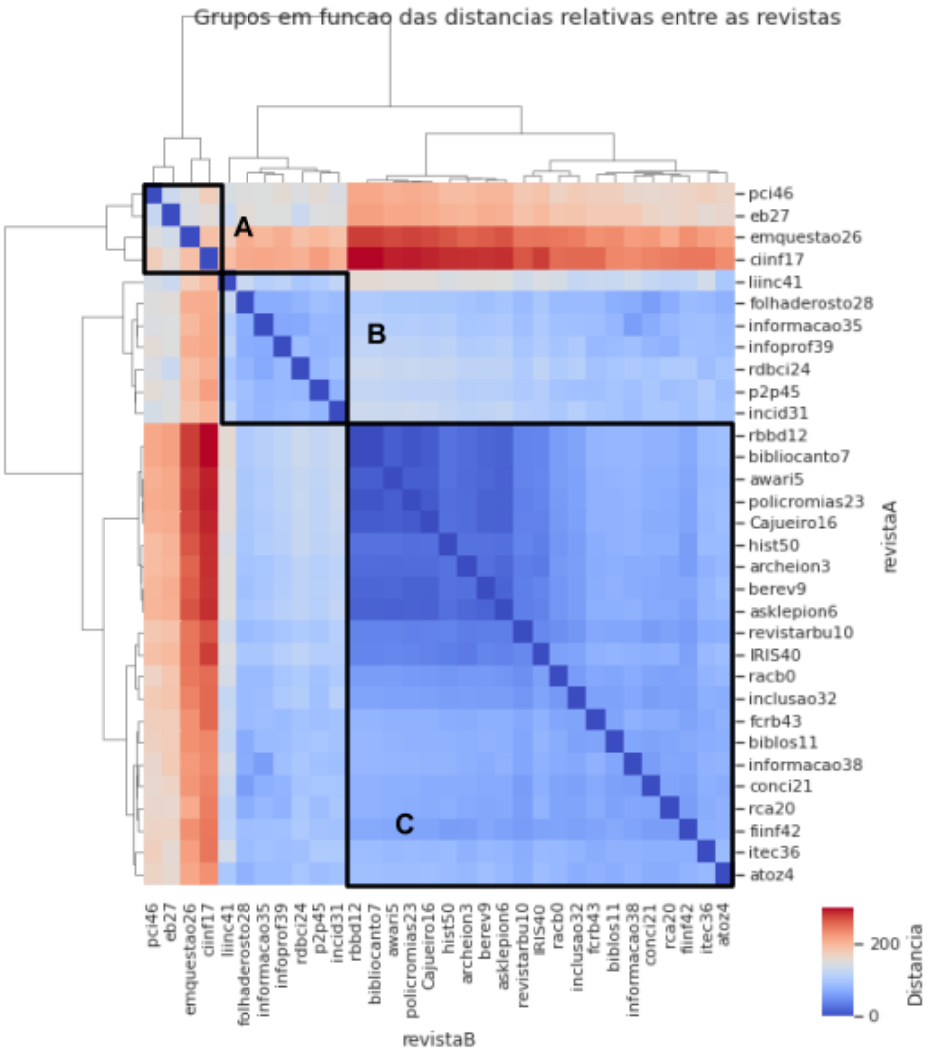


Fonte: Elaborado pelo autor.

Encontrou-se uma correlação negativa (-0.8) entre a quantidade de artigos e a distância ao tesauro, assim como entre o tamanho do grafo gerado e a distância, de forma que quantos mais artigos presentes na amostra da revista maior é o grafo e menor a sua distância para o grafo do tesauro.

Após o cálculo da distância ao tesauro, foram calculadas as distância entre as revistas, com o seguinte agrupamento. Diante o dendrograma referente aos agrupamentos identificou-se que um número de 3 grupos podem descrever as relações entre as coberturas temáticas das revistas. Por meio da "clusterização" hierárquica e aglomerativa das revistas, considerando-se a distância a cada outra revista com um atributo, foi possível associar cada revista a um determinado grupo, resultando no diagrama da figura 3. De modo a facilitar a visualização, foram adotadas identificadores para cada revista, que podem ser consultados Anexo 1, assim como os respectivos grupos de cada revista.

Figura 3 - Mapa de calor com agrupamentos das revistas em função de sua distância relativa



Fonte:

Elaborado pelo autor.

Os grupos de revistas, identificados A, B e C, podem ser caracterizados da seguinte forma:

- grupo A, menor quantidade de revistas, pouca similaridade entre as próprias revistas do grupo, média similaridade com o grupo B e baixa similaridade com o grupo C;
- grupo B, média quantidade de revistas, média similaridade entre as próprias revistas do grupo, baixa similaridade com revistas do grupo A, e maior similaridade entre as revistas grupo C;
- grupo C, grande quantidade de revistas, alta similaridade entre as próprias revistas do grupo, media similaridade com revistas do grupo B, e baixa similaridade com o grupo A;

Diante a caracterização dos grupos de revistas em função de sua aderência temática, podem ser apresentadas algumas considerações finais e conclusão.

4 CONCLUSÕES

Contata-se que há pouca bibliografia relacionada a uma métrica sobre cobertura temática, de tal forma que este trabalho busca contribuir para as discussões sobre o assunto.

O índice de cada revista pode ser calculado de forma absoluta em relação ao tesauro e também de forma relativa entre revistas.

Diante os resultados apresentados na figura 3 é possível afirmar que, diante a cobertura temática inferida a partir da técnica de extração de termos utilizada, as revistas Ciência da Informação, Em Questão, Encontros Bibli e Perspectivas em Ciência da Informação, pertencentes ao grupo A, formam um conjunto heterogêneo, com coberturas temáticas distintas tanto entre si quanto em relação a outras revistas. Esse grupo contém também a maior parte de revistas com menores distâncias em relação ao tesauro. Por sua vez, diante da perspectiva da cobertura temática, as revistas do grupo C são mais semelhantes entre si, formando um grupo de revistas com temáticas comuns.

Considerando-se o tesauro uma ferramenta que fornece uma linguagem documental que visa otimizar a indexação e recuperação de documentos, a proximidade do grafo das revistas com o grafo do tesauro indica, de forma indireta, um maior potencial de recuperação dos documentos em relação a revistas com menor semelhança.

Diante o não uso das palavras-chave fornecidas pelo usuário, mas sim sua extração automática, para o levantamento da cobertura temática, não cabe dizer que uma melhoraria da métrica de cobertura temática seria obtida por melhores práticas de catalogação no processo editorial.

Quanto à correlação entre a quantidade de artigos e a cobertura temática, esta não é perfeita, havendo revistas que conseguem abranger uma cobertura maior do que outras mesmo tendo menor quantidade de artigos. Assim, como trabalho futuro sugere-se investigar também a relação da quantidade de artigos necessária para contemplar determinada quantidade de tópicos do tesauro, identificando-se revistas quem tem maior consistência temática do que outras.

Finalmente, é apresentada uma ferramenta para levantamento de uma métrica que permite comparar diferentes periódicos em relação a uma cobertura temática de temas definidos em um tesauro da área. Os algoritmos utilizados no estudo foram disponibilizados na forma de Jupyter Notebook (BRITO, 2022) permitindo sua repetição da análise assim como sua aplicação com outras revistas ou tesauros de outras áreas do conhecimento.

REFERÊNCIAS

LIDDY, E. D. Natural Language Processing. *In:* ENCYCLOPEDIA of Library and Information Science. 2. ed. New York: Marcel Decker, 2001.

PINHEIRO, L. V. R.; FERREZ, H. D. **Tesauro Brasileiro de Ciência da Informação**. Rio de Janeiro; Brasília: Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), 2014.

AIZAWA, A. An information-theoretic perspective of tf–idf measures. **Information Processing & Management**, London, v. 39, n. 1, p. 45–65, jan. 2003.

ALVAREZ-OSSORIO, J. R. P. Cobertura temática y procedencia institucional de los artículos publicados en la Revista Española de Documentación Científica en sus veinte años de existencia. **Revista Española de Documentación Científica**, [*s.l.*], v. 20, n. 3, p. 290–298, 30 set. 1997.

BRITO, Ronnie F. **Jupyter Notebook para levantamento de cobertura temática**. [*S.l.*]: Zenodo. 2022. Disponível em: https://doi.org/10.5281/zenodo.6249967. Acesso em: 20 abr. 2022.

BRITO, Ronnie F. Grafos **de cobertura temática de revistas da área de Ciência da Informação [Data set]**. [*S.l.*]: Zenodo. 2022. Disponível em: https://doi.org/10.5281/zenodo.6533926. Acesso em: 20 abr. 2022.

CORONINI, R.; MANGEMATIN, V. From individual scientific visibility to collective competencies: The example of an academic department in the social sciences. **Scientometrics**, Budapest, v. 45, n. 1, p. 55–80, maio 1999.

CURIEL LORENZO, S.; HERNÁNDEZ ACOSTA, L. **Análisis de la cobertura temática de la revista electrónica "Avanzada Científica. Journal article (Unpaginated, i.e., html or from an issue unpaginated)**. Disponível em: http://eprints.rclis.org/16060/. Acesso em: 15 fev. 2022.

DE MOYA-ANEGÓN, F. *et al*. Coverage analysis of Scopus: A journal metric approach. **Scientometrics**, Budapest, v. 73, n. 1, p. 53–78, out. 2007.

GAO, X. et al. A survey of graph edit distance. Pattern Analysis and Applications, v. 13, n. 1, p. 113–129, fev. 2010.

KORENČIĆ, D. *et al*. A Topic Coverage Approach to Evaluation of Topic Models. IEEE Access, [*s.l.*], v. 9, p. 123280–123312, 2021.

MULLER, C. C.; OLIVEIRA, K. S. **Repositório Institucional da Enap: um processo de construção coletiva do conhecimento**. Disponivel em: http://www.enap.gov.br/index.php?option=content&task=view&id=258. Acesso em: 15 maio 2015.

VUOTTO, A.; FERNANDEZ, G.; BOGETTI, C. Aplicación del factor TF-IDF en el análisis semántico de una colección documental. **Biblios**: Revista de Bibliotecología y Ciencias de la Información, v.60, 2015.

ZACCA-GONZÁLEZ, G. et al. Bibliometric analysis of regional Latin America's scientific output in Public Health through SCImago Journal & Country Rank. **BMC Public Health**, London, v. 14, n. 1, p. 632, 21 jun. 2014.

ANEXOS

A - Quadro de revistas analisadas

Revista	Identificador	Qtde de Artigos	Tamanho do Grafo	Distância Tesauro	Grupo
Revista ACB	racb0	675	120	1677	C
Revista Acervo	revistaacervo1	nc	nc	np	
Revista Agora	agora2	nc	nc	np	
Revista Archeion	archeion3	95	26	np	
Revista AtoZ	atoz4	188	65	1733	C
Revista AWARI	awari5	14	4	np	
Revista Asklepion: Informação em Saúde	asklepion6	20	2	np	
Revista Bibliocanto	bibliocanto7	nc	nc	np	
Revista Biblionline	biblio8	100	39		
Revista Biblioteca Escola em Revista	berev9	93	14	np	
Revista Bibliotecas Universitárias: pesquisas ...	revistarbu10	63	31	1761	C
Revista Biblos	biblos11	602	116	1746	C
Revista Brasileira de Biblioteconomia e Docume...	rbbd12	nc	nc	np	
Revista Brasileira de Educação em Ciência da I...	rebecin13	nc	nc	np	
Revista Brazilian Journal of Information Science	bjis14	nc	nc	np	
Revista Cadernos de Informação Jurídica	cajur15	nc	nc	np	
Revista Cajueiro	Cajueiro16	50	4	np	
Revista Ciência da Informação	ciinf17	951	213	1752	A
Revista Ciência da Informação em Revista	cir18	nc	nc	np	
Revista Comunicação e Informação	ci19	nc	nc	np	
Revista Conhecimento em Ação	rca20	50	33	1758	C
Revista Convergências em Ciência da Informação	conci21	89	36	1752	C
Revista de Biblioteconomia de Brasília	rbbsb22	nc	nc	np	
Revista de Estudos do Discurso Imagem e Som - ...	policromias23	39	4	np	
Revista Digital de Biblioteconomia & .	rdbci24	537	150	1720	B
Revista Eletrônica Informação e Cognição	reic25	nc	nc	np	
Revista Em Questão	emquestao26	740	168	1722	A
Revista Encontros Bibli	eb27	476	147	1682	A
Revista Folha de Rosto	folhaderosto28	213	70	1742	B
Revista Fontes Documentais	fontesdocumentais29	nc	nc	np	
Revista Ibero-americana de Ciência da Informação	rici30	nc	nc	np	
Revista InCID	incid31	339	117	1740	B
Revista Inclusão Social	inclusao32	354	57	1743	C
Revista Informação & Sociedade: Estudos	ies33	77	13	np	
Revista Informação Arquivística	informacaoarquivistica34	nc	nc	np	
Revista Informação e Informação	informacao35	169	93	1746	B
Revista Informação e Tecnologia	itec36	98	56		
Revista Informação em Pauta	informacaoempauta37	192	75		
Revista Informação na Sociedade Contemporânea	informacao38	169	93	1745	
Revista Informação@Profissões	infoprof39	181	75	1738	
Revista IRIS	IRIS40	37	13	np	
Revista Liinc	liinc41	582	129	1707	B
Revista Logeion	fiinf42	155	29	1745	
Revista Memória e Informação	fcrb43	90	29	1742	
Revista Multiplos Olhares em Ciência da Inform...	moci44	nc	nc	np	
Revista P2P e INOVAÇÃO	p2p45	253	55	1724	B
Revista Perspectivas em Ciência da Informação	pci46	972	212	1712	A
Revista Perspectivas em Gestão & Conhecimento	pgc47	100	28	np	
Revista Pesquisa Brasileira em Ciência da Info...	pbcib48	100	1	np	
Revista Ponto de Acesso	revistaici49	nc	nc	np	
Revista Revista do Departamento de Bibliotecon...	hist50	248	16	np	
Revista Transinformação	transinfo51	nc	nc	np	
nc=Não Coletada np = Não processada					