

CIENTOMETRIA E MINERAÇÃO DE DADOS: um relato de uma experiência com recursos da Web de Dados

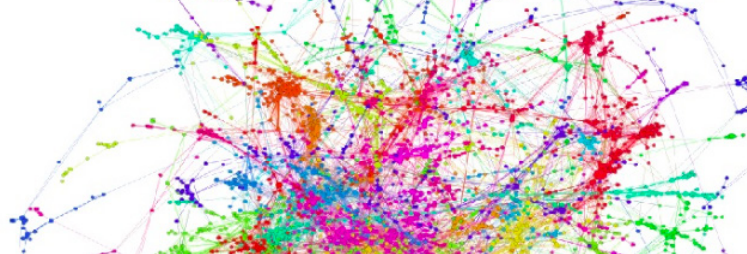
Paulo R. V. do Carmo
Universidade Estadual do Centro-Oeste, Guarapuava, PR, Brasil
pauloviviurka4@gmail.com

Sandro Rautenberg
Universidade Estadual do Centro-Oeste, Guarapuava, PR, Brasil
srautenberg@unicentro.br

1 INTRODUÇÃO

Esta pesquisa permeia três construtos: Cientometria, Web de Dados e Mineração de Dados. A Cientometria é a base conceitual para os estudos que utilizam os registros de documentos científicos (resumos, artigos, livros, patentes, dentre outros) como dados primários para responder às questões que mensuram a informação científica/tecnológica (STOCK; STOCK, 2015). Instrumentalmente, pressupõe-se que os conjuntos de dados primários são capturados a partir da Web de Dados. Neste enredo, a Web de Dados desempenha o papel de uma fonte global de recursos de informação, em que dados são inter-relacionados e disponibilizados abertamente para o reuso (W3C, 2017), inclusive para com as pesquisas cientométricas. Por exemplo, os conjuntos de dados sobre autores de comunicações científicas, periódicos, fatores de impacto, são compartilhados abertamente na Web de Dados (RAUTENBERG et al., 2017a; RAUTENBERG et al., 2017b), permitindo o desenvolvimento de pesquisas no campo da Ciência da Informação.

Entretanto, em face do elevado volume de dados disponível na Web de Dados, explorar os recursos disponibilizados em busca de informação relevante não é uma atividade trivial. Nesta perspectiva, Wolfram (2017) pontua que pesquisadores da Ciência da Informação podem



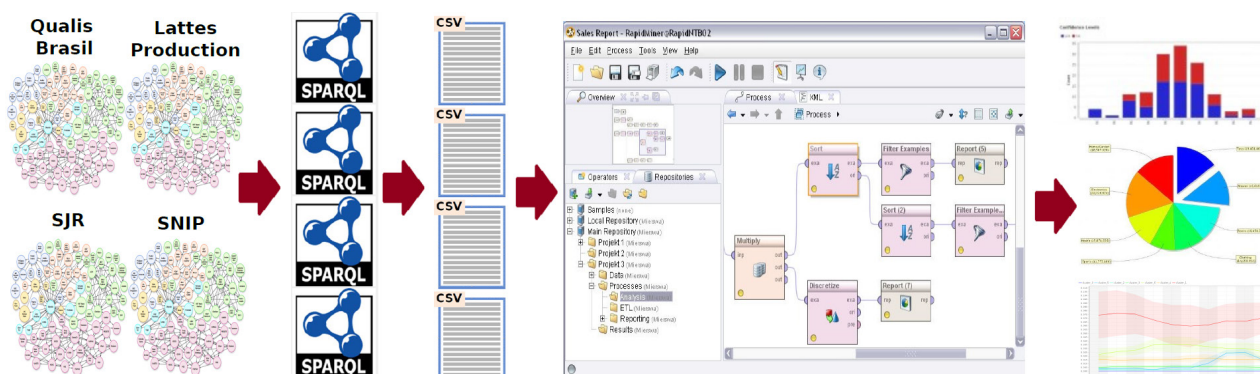
utilizar técnicas da Mineração de Dados. A Mineração de Dados envolve a aplicação de processos computacionais sobre bases de dados para, automaticamente, obter informações relevantes e que, *a priori*, são desconhecidas (FLYNN, 2002, p. 233).

Diante os construtos resumidamente apresentados e considerando o domínio de uma universidade pública brasileira, este trabalho investiga algumas tecnologias computacionais aderentes à obtenção automatizada de informações científicas relevantes a partir de dados oriundos da Web de Dados.

2 MATERIAIS E MÉTODOS

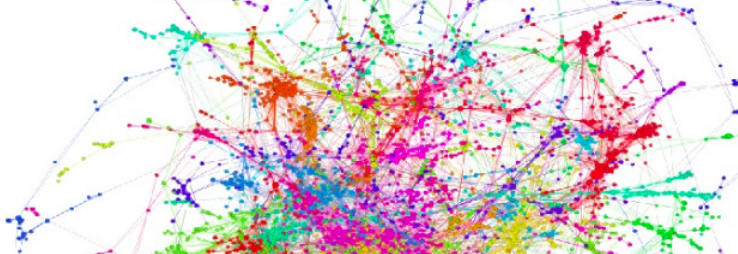
Na Web de Dados, quatro conjuntos de dados primários são disponibilizados para pesquisas cientométricas, conforme metodologicamente discutido em Rautenberg et al. (2017a) e Rautenberg et al. (2017b). Neste sentido, são utilizados: **i) QualisBrasil** – o histórico do índice Qualis (período 2005-2016); **ii) SJR** – o histórico do fator de impacto SJR (período 2005-2016); **iii) SNIP** – o histórico do fator de impacto SNIP (período 2005-2016); e **iv) LattesProduction** – o conjunto dos registros de artigos publicados em revistas de uma universidade pública brasileira (período 2005-2016).

FIGURA 1 - PROCEDIMENTO METODOLÓGICO ESTABELECIDO



Fonte: Dados da pesquisa.

O procedimento metodológico adotado é representado na Figura 1. Inicialmente, os conjuntos de dados primários são capturados da Web

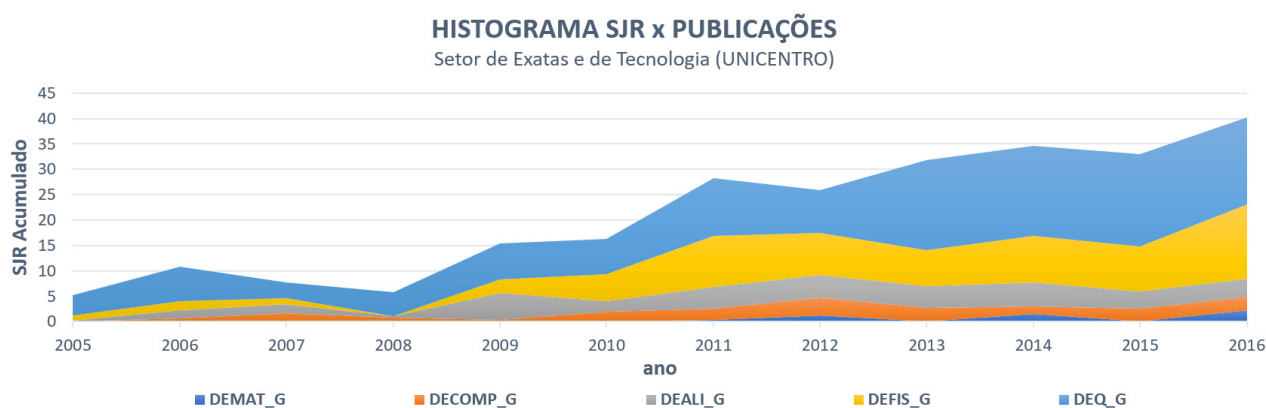


de Dados mediante consultas SPARQL. Os dados devem ser automaticamente armazenados em arquivos CSV (*Comma Separated Value*). Os arquivos CSV são importados em um projeto de análise de dados na ferramenta Rapidminer^I. Uma vez realizada a importação, os dados são relacionados (considerando o ano e o ISSN da revista dos artigos publicados). Por fim, os cenários de exploração dos dados devem ser estabelecidos para obter informações cientométricas relevantes.

3 RESULTADOS PARCIAIS

Perante o procedimento metodológico adotado, têm-se como resultados parciais alcançados: **i)** as consultas SPARQL foram desenvolvidas e executadas, adquirindo os dados primários da Web de Dados; **ii)** os dados foram importados para um projeto de mineração de dados na ferramenta Rapidminer; e **iii)** em um *workflow*, os conjuntos de dados estão relacionados mediante os anos de avaliação e os ISSNs dos periódicos.

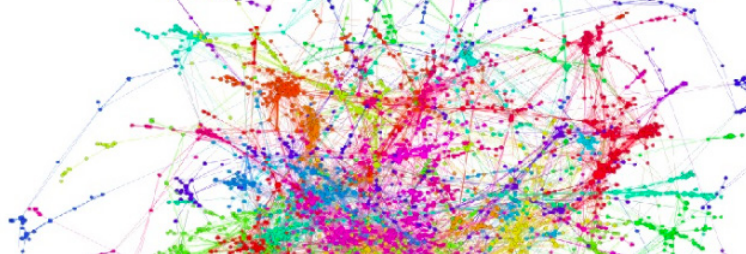
FIGURA 2 - EXEMPLO DE CENÁRIO DE EXPLORAÇÃO INFORMACIONAL



Fonte: Dados da pesquisa.

Como esforço atual, alguns estudos dirigidos são realizados, principalmente, privilegiando o aprendizado da utilização da Rapidminer. Neste sentido, como um exemplo de exploração, na Figura 2 é repre-

^I Rapidminer é uma plataforma de software voltada a cientistas de dados. Em aplicações de mineração de dados, é usada desde o tratamento dos dados primários, até a identificação de padrões de comportamento nos dados. Disponível em: <<https://rapidminer.com/pricing/>>. Acesso em: 19 dez. 2017.



sentado um estudo preliminar. Neste cenário objetiva-se mensurar a evolução da qualidade das publicações realizadas pelos professores de um centro de conhecimento em uma universidade pública, organizados segundo seus departamentos de origem. Para tanto, os fatores de impacto atrelados às publicações são sumarizados, anualmente. No exemplo, destaca-se a evolução da qualidade de publicações do subgrupo dos professores do Departamento de Química da Universidade Estadual do Centro-Oeste – UNICENTRO (o escore SJR acumulado representado em azul – DEQ_G). Considerando a área plotada, admite-se uma tendência de crescimento na disseminação das pesquisas. Ressalta-se que a referida tendência é enfatizada a partir do ano de 2008, quando da instauração do curso de Pós-Graduação em nível de Doutorado em Química em conjunto às Universidade Estadual de Londrina (UEL) e Universidade Estadual de Ponta-Grossa (UEPG).

4 CONSIDERAÇÕES FINAIS

Atualmente, a Web de Dados constitui-se em uma fonte de informação de escala global. Para explorar a capacidade informacional desta infraestrutura, pode-se recorrer às técnicas e ferramentas de Mineração de Dados para obter informação relevante nos mais variados domínios. Considerando essas assertivas, neste trabalho pressupôs-se que a Web de Dados é uma importante fonte de dados para pesquisas no campo da Ciência da Informação.

Neste sentido, pontualmente, verificou-se o potencial supracitado na formalização de uma abordagem metodológica para realização de pesquisas cientométricas mais robustas com recursos provenientes da Web de Dados. Entretanto, dada a infância deste trabalho, novos ensaios cientométricos necessitam ser conduzidos para refinar a abordagem proposta ou obter informações cientométricas relevantes. Em face a isso, como trabalhos futuros, pretende-se traçar novos cenários junto à universidade em investigação para contextualizar e fomentar o conhecimento científico institucional.



AGRADECIMENTOS

Os autores agradecem à Fundação Araucária pela bolsa de Iniciação Científica concedida (Programa Institucional De Iniciação Científica – PROIC/UNICENTO).

REFERÊNCIAS

FLYNN, R. R. **Computer Science: Foundation; Ideas and People**. New York: Macmillan, 2002.

RAUTENBERG, S.; MOTYL, S.; BURDA, A.; SILVERIO, A.; MOURA, F. M. Dados Abertos Conectados e Gestão do Conhecimento: estudos de caso cientométricos em uma universidade brasileira. **Perspectivas em Ciência da Informação**, v. 22, p. 116-142, 2017a.

RAUTENBERG, S.; SOUZA, L.; HAUAGGE, J.; HILD, T.; MICHELON, G.; BURDA, A. Representando índices cientométricos como dados abertos conectados. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017, Marília. **Anais ...** Marília-SP: PPGCI, UNESP, 2017b.

STOCK, W. G.; STOCK, M. **Handbook of Information Science**. Berlin: Walter de Gruyter GmbH, 2015.

W3C. **Linked Data**. Disponível em: <<https://www.w3.org/standards/semanticweb/data>>. Acesso em: 19 dez. 2017.

WOLFRAM, D. Bibliometrics Research in the Era of Big Data: Challenges and Opportunities. In: MUGNAINI, R.; FUJINO, A.; KOBASHI, N. Y. (Ed.). **Bibliometria e cientometria no Brasil: infraestrutura para avaliação da pesquisa científica na era do Big Data**. São Paulo: ECA/USP, 2017.