

VISUALIZAÇÃO DA COLABORAÇÃO CIENTÍFICA ENTRE PESQUISADORES A PARTIR DE METADADOS DA PLATAFORMA LATTES

Vinícius Rafael Micali Soares

Mestre em Ciência da Informação
Analista de TI na Universidade Federal de São Carlos (UFSCar)
E-mail: viniciusmicali@ufscar.br

Luís Gustavo Maschietto

Mestre em Ciência, Tecnologia e Sociedade
Professor no Centro Paula Souza
E-mail: luis.maschietto@etec.sp.gov.br

Mesailde Souza de Oliveira Matias

Técnica em TI na Universidade Federal de São Carlos (UFSCar)
E-mail: mesailde@ufscar.br

Paulo Matias

Professor do Magistério Superior na Universidade Federal de São Carlos (UFSCar)
E-mail: matias@ufscar.br

Roniberto Morato do Amaral

Professor do Magistério Superior na Universidade Federal de São Carlos (UFSCar)
E-mail: roniberto@ufscar.br

Leandro Innocentini Lopes de Faria

Professor do Magistério Superior na Universidade Federal de São Carlos (UFSCar)
E-mail: leandro@ufscar.br

INTRODUÇÃO

Várias iniciativas têm procurado explorar os metadados da Plataforma Lattes para a produção de indicadores de ciência e tecnologia, superando diversos desafios. Uma das ferramentas é o scriptLattes, um script robô escrito em linguagem de programação Python, que extrai informações dos currículos formatados em HTML (Hyper Text Markup Language) e disponibiliza publicamente, no site da Plataforma Lattes. (MENA-CHALCO; CESAR JÚNIOR, 2013)

Com o uso do DOI (Digital Object Identifier) desde setembro de 2008 pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), foram minimizados os

problemas referentes à qualidade dos metadados das publicações, principalmente artigos de periódicos, uma vez que basta ao pesquisador indicar o DOI do artigo a ser cadastrado no Currículo Lattes para que os dados sejam obtidos diretamente do sistema DOI (mantido pela IDF - International DOI Foundation), o que ajuda na identificação de coautorias. No entanto, não há garantias de que o pesquisador irá preencher o número do DOI dos artigos constantes em seu currículo. Para publicações sem DOI, as informações do artigo são informadas manualmente pelos pesquisadores, o que pode acarretar diversos erros, como nome de pesquisador informado de várias maneiras, publicações com títulos diferentes em currículos distintos, ausência de dados, etc. Além disso, há o desafio de deduplicar publicações extraídas de forma automatizada, uma vez que um mesmo artigo publicado em coautoria entre autores está nos currículos de cada autor e acaba duplicado no conjunto de dados.

A ferramenta *synclattes* desenvolvida por Matias (2015), além de abordar os problemas mencionados de forma mais abrangente que o *scriptLattes*, utiliza uma fonte de dados de acesso mais direto – o currículo em formato XML (Extensible Markup Language), que torna desnecessária a etapa de pré-processamento do HTML empregada pelo *scriptLattes*. Assim como o currículo em HTML, o currículo em XML está disponível publicamente com o preenchimento de um CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart), mas também pode ser obtido automaticamente, por meio de convênio que o CNPq fornece às instituições. Para a contagem de metadados dos currículos e geração de visualizações diversas podem ser usados programas bibliométricos, aplicações estatísticas ou até mesmo planilhas eletrônicas. Resta, porém, o desafio de gerar visualizações gráficas na Web usando dados já padronizados a partir dos metadados extraídos da Plataforma Lattes pela ferramenta automatizada *synclattes*, o que compreende o objetivo deste trabalho, que buscou representar de forma visual a colaboração científica entre pesquisadores.

A área de visualização de informação pode contribuir para a compreensão dos resultados das iniciativas científicas e tecnológicas em que os pesquisadores brasileiros estão envolvidos, uma vez que, com a sobrecarga de informações, o estudo de como representá-las de uma forma visual, que facilite sua compreensão, tem sido uma das principais preocupações dos estudos dessa área. (ALVES, 2015) De acordo com Nascimento e Ferreira (2005), por meio de uma visualização gráfica é possível sintetizar grande quantidade de informações.

Devido ao impacto no desenvolvimento científico, é crescente o interesse em estudos métricos sobre colaborações científicas, sobretudo as internacionais, essas que proporcionam alta visibilidade às instituições de ensino superior nacionais envolvidas, incitando a internacionalização dos Programas de Pós-Graduação. (JUSTINO, 2019) Por causa da importância e respectiva complexidade, torna-se fundamental a existência de ferramentas que mensurem a colaboração entre pesquisadores, dado que os indicadores

de produção científica são essenciais para a compreensão dos impactos das políticas e dinâmicas sociais envolvendo ciência e tecnologia.

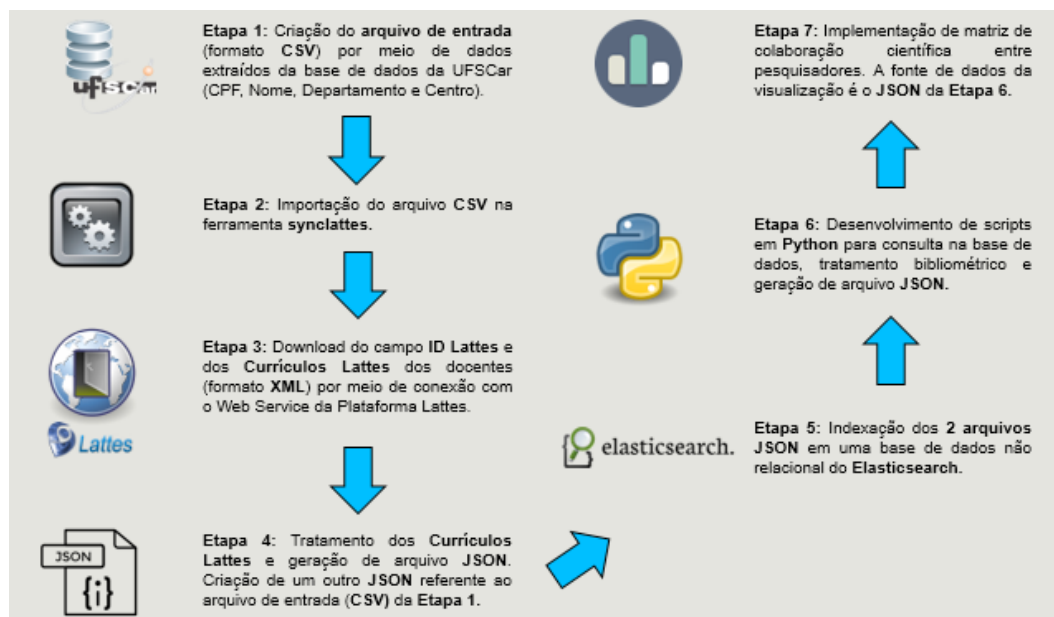
O software desenvolvido como parte desta pesquisa trabalha com arquivos no formato JSON e torna desnecessário o uso de outros softwares (por ex., Microsoft Excel) para importação de arquivos texto e posterior geração de visualizações.

METODOLOGIA

Levando em consideração Gil (2008), esta pesquisa é de natureza aplicada e exploratória, sendo que sua aplicação ocorreu na unidade caso Universidade Federal de São Carlos (UFSCar), fundada em 1968 e atualmente formada pelos *campi* São Carlos, Araras, Sorocaba e Lagoa do Sino (em Buri), constituindo-se uma instituição pública de relevância no cenário nacional e internacional, destacando-se pelo alto nível de qualificação de seu corpo docente, que em sua maioria desenvolve atividades de ensino, pesquisa e extensão em regime de dedicação exclusiva.

A partir de arquivos no formato JSON (JavaScript Object Notation) que foram exportados pela ferramenta *synclattes*, foi criado um procedimento para tratamento bibliométrico com a utilização da linguagem Python e de uma base de dados não relacional do servidor de buscas Elasticsearch, o que permitiu a geração de um arquivo JSON para a construção da visualização. A Figura 1 apresenta uma síntese do método utilizado no desenvolvimento desta pesquisa, envolvendo a realização de sete macro etapas.

Figura 1: Fluxo das atividades



Fonte: adaptada de Soares (2019).

Das etapas 1 a 4, foi realizada a extração da produção científica da UFSCar por meio da ferramenta synclattes, a qual gerou dois arquivos no formato JSON, um com informações referentes aos dados pessoais dos autores (obtidos na base de dados da UFSCar) e outro contendo informações provenientes dos currículos Lattes (metadados de artigos publicados em eventos e em periódicos). Em relação à estrutura desses arquivos, há um identificador de autor que é comum entre eles, denominado “Idlattes” no arquivo de autores e “IdCNPq” no arquivo de dados de artigos. Com esse identificador é possível cruzar os dados dos dois arquivos, o que permite encontrar por meio de consultas toda a produção científica de um autor ou mesmo a sua rede colaborativa. Entre as etapas 5 e 7, os arquivos JSON foram indexados em uma base de dados não relacional, uma consulta foi realizada para contar a colaboração científica entre os pesquisadores da UFSCar e um novo arquivo JSON foi criado, o que possibilitou a disponibilização de uma visualização de forma matricial.

No presente trabalho foi desenvolvida uma matriz que representa a colaboração científica entre pesquisadores. Em relação à amostra (recorte), optou-se por considerar os 50 pesquisadores da UFSCar que mais publicaram artigos em eventos e em periódicos. Santos e Mugnaini (2019) argumentam que o periódico é um veículo de publicação usado pela comunidade científica que geralmente é focado em uma determinada área do conhecimento, em que reflete as características dessa área em específico.

RESULTADOS

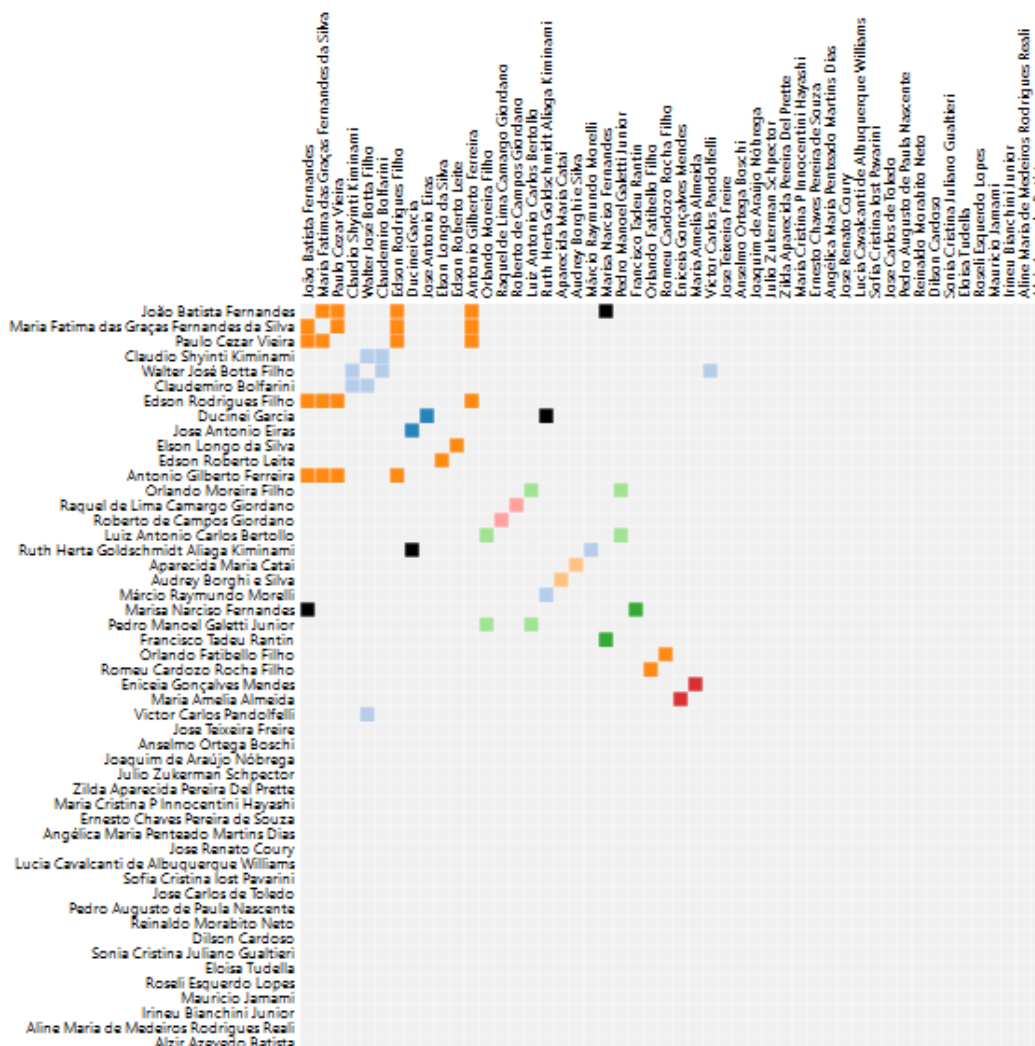
Para tornar possível a realização de consultas parametrizadas a partir da definição de variáveis de entrada (filtros), os arquivos fornecidos pela synclattes foram indexados na base de dados não relacional do Elasticsearch, que é uma ferramenta de busca *open-source* (código aberto) que possui diversos recursos computacionais com a finalidade de disponibilizar em tempo real os documentos armazenados, tornando-os indexados e pesquisáveis. As requisições à base de dados do Elasticsearch podem ser realizadas por linha de comando ou por meio de um servidor de aplicação local que disponibiliza uma API, que é uma interface de programação de aplicativos. (GORMLEY; TONG, 2015) A escolha pelo Elasticsearch deve-se a praticidade para trabalhar com as informações textuais contidas nos arquivos JSON utilizados neste trabalho (1.842 registros de informações pessoais de autores e aproximadamente 90.000 registros de publicações).

Após a importação no Elasticsearch dos dois arquivos JSON exportados pela synclattes, foram implementados scripts em Python para consulta e tratamento bibliométrico dos currículos dos pesquisadores contidos na base de dados indexada no Elasticsearch. Como resultado, foi criado um arquivo JSON contendo os 50 pesquisadores da UFSCar que mais publicaram artigos em eventos e em periódicos, bem como a respectiva colaboração científica entre esses pesquisadores. A partir desse arquivo JSON e com base

nos conceitos da área de visualização de informação, foi desenvolvida uma matriz que representa de forma visual a coautoria entre pesquisadores.

Para a implementação dessa visualização foi utilizada a biblioteca gráfica de código livre D3 (Data Driven Documents), que foi criada em linguagem JavaScript e tem por objetivo a manipulação de documentos com paradigma de programação centrado em dados. A escolha por essa tecnologia foi baseada na complexidade da visualização, o que vai ao encontro do que vem sendo mostrado no portal VisPublica (2020), que apresenta exemplos de utilização de várias ferramentas e técnicas de visualização de informação. A Figura 2 apresenta a matriz de colaboração científica entre os 50 pesquisadores da UFS-Car que mais publicaram artigos em eventos e em periódicos.

Figura 2: Matriz de colaboração científica entre pesquisadores



Fonte: adaptada de Soares (2019).

Os pesquisadores são mostrados nos vértices e as interseções entre eles nas células da matriz. As cores representam os departamentos que os pesquisadores trabalham. Quando pesquisadores de departamentos diferentes possuem relação (publicação em coautoria), a célula de intersecção entre eles é preenchida pela cor preta. Para cada pesquisador é realizada a contagem dos artigos que ele publicou com os outros 49 pesquisadores mais produtivos da UFSCar, sendo os nomes dispostos por frequência (pesquisadores que mais publicaram em coautoria aparecem antes).

Como pode ser observado na Figura 2, os três pesquisadores que mais publicaram em coautoria são do mesmo departamento. Porém, o primeiro deles possui publicação com alguém de outro departamento. É importante destacar que dos 50 pesquisadores da UFSCar que mais publicaram artigos em eventos e em periódicos, vários deles não têm publicação em coautoria com algum outro pesquisador contido nesse recorte, o que pode ser facilmente visto por meio dessa visualização em forma de matriz, bastando percorrer uma linha ou uma coluna em que há somente células na cor cinza.

CONSIDERAÇÕES FINAIS

O processo de construção do conhecimento e da ciência está relacionado ao potencial de publicação e divulgação dos resultados de pesquisas. Apesar da existência de diversos estudos sobre elaboração de indicadores bibliométricos, há a necessidade de se desenvolver soluções tecnológicas que possibilitem a visualização desses indicadores.

A Plataforma Lattes apresenta-se como uma relevante fonte de dados para a elaboração de indicadores de ciência e tecnologia e este trabalho mostrou a viabilidade da implementação de visualização para a representação gráfica do indicador de produção científica, a partir de metadados extraídos pela ferramenta sync lattes. Futuras pesquisas poderiam ampliar a amostra analisada de pesquisadores, e ainda, comparar os resultados obtidos com outros softwares que trabalham com a Plataforma Lattes para a geração de indicadores científicos.

Um desafio visto em softwares bibliométricos é o preparo dos dados que alimentarão a ferramenta que gera a visualização, uma vez que cada sistema possui padrões de entrada específicos. A ferramenta desenvolvida nesta pesquisa não tem esse problema, pois ela trabalha com a notação JSON, que é um formato aberto que permite facilmente a troca de dados entre sistemas.

Conclui-se que é possível prover visualizações que favoreçam a compreensão das estruturas e dinâmicas da ciência brasileira, podendo, assim, contribuir para estudos a respeito da avaliação das políticas públicas de ciência e tecnologia no Brasil.

REFERÊNCIAS

ALVES, M. C. *Visualização de informação para simplificar o entendimento de indicadores sobre avaliação da ciência e tecnologia*. 2015. Dissertação (Mestrado em Ciência, Tecnologia e Sociedade) - Programa de Pós Graduação em Ciência, Tecnologia e Sociedade, Universidade Federal de São Carlos, São Carlos, 2015.

GIL, A. C. *Como elaborar projetos de pesquisa*. 6. ed. São Paulo: Atlas, 2008.

GORMLEY C.; TONG Z. *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine*. Sebastopol: O'Reilly Media, 2015.

JUSTINO, T. S. *Análise da Colaboração Científica dos Programas de Pós-Graduação em Ciência da Informação Brasileiros*. 2019. Dissertação (Mestrado em Ciência da Informação) - Programa de Pós Graduação em Ciência da Informação, Universidade Federal de São Carlos, São Carlos, 2019.

MATIAS, M. S. O. *Base referencial para o povoamento de repositórios institucionais: coleta automatizada de metadados da Plataforma Lattes*. 2015. Dissertação (Mestrado em Gestão de Organizações e Sistemas Públicos) - Universidade Federal de São Carlos, São Carlos, 2015.

MENA-CHALCO, J. P.; CESAR JÚNIOR, R. M. Prospecção de dados acadêmicos de currículos Lattes através de scriptLattes. In: HAYASHI, M. C. P. I.; LETA, J. (org.). *Bibliometria e Cientometria: reflexões teóricas e interfaces*. São Carlos: Pedro & João Editores, 2013. p. 109-128.

NASCIMENTO, H. A.; FERREIRA, C. B. R. Visualização de Informações: uma abordagem prática. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 25., 2005, São Leopoldo. *Anais [...]*. São Leopoldo: UNISINOS, 2005. p. 1262-1312.

SANTOS, S. M.; MUGNAINI, R. Comunicação científica em periódicos e a internacionalização das ciências brasileiras. In: CARNEIRO, F. F. B.; FERREIRA NETO, A.; SANTOS, W. (org.). *A comunicação científica em periódicos*. Curitiba: Appris, 2019. p. 73-93.

SOARES, V. R. M. *Interface interativa para visualização de indicadores de produção científica a partir de dados bibliométricos processados da Plataforma Lattes*. 2019. Dissertação (Mestrado em Ciência da Informação) - Programa de Pós Graduação em Ciência da Informação, Universidade Federal de São Carlos, São Carlos, 2019.

VISPUBLICA. Modelo de visualização de dados públicos. *Vispublica*, [2020], 2020. Disponível em: <http://vispublica.gov.br>. Acesso em: 25 maio 2020.

