UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA, TECNOLOGIA E SOCIEDADE

ANÁLISE DA ENDOGAMIA DAS COAUTORIAS CIENTÍFICAS DE PESQUISADORES DA UFSCAR CREDENCIADOS EM PROGRAMAS DE PÓS-GRADUAÇÃO

VINÍCIUS RAFAEL MICALI SOARES

SÃO CARLOS - SP
Março/2026

VINÍCIUS RAFAEL MICALI SOARES

ANÁLISE DA ENDOGAMIA DAS COAUTORIAS CIENTÍFICAS DE PESQUISADORES DA UFSCAR CREDENCIADOS EM PROGRAMAS DE PÓS-GRADUAÇÃO

Tese apresentada ao Programa de Pós-Graduação em Ciência, Tecnologia e Sociedade, do Centro de Educação e Ciências Humanas, da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Doutor em Ciência, Tecnologia e Sociedade.

Orientador: Prof. Dr. Leandro Innocentini Lopes de Faria

SÃO CARLOS - SP
Março/2026

UFSCar

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Educação e Ciências Humanas

Programa de Pós-Graduação em Ciência, Tecnologia e Sociedade

Folha de Aprovação

Defesa de Tese de Doutorado do candidato Vinícius Rafael Micali Soares, realizada em 13/03/2026.

Comissão Julgadora:

Prof. Dr. Leandro Innocentini Lopes de Faria (UFSCar)

Prof. Dr. Roniberto Morato do Amaral (UFSCar)

Profa. Dra. Luciana de Souza Gracioso (UFSCar)

Prof. Dr. Fernando de Assis Rodrigues (UFPA)

Prof. Dr. Jesús Pascual Mena Chalco (UFABC)

O Relatório de Defesa produzido pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação.

PUBLICAÇÕES RELACIONADAS À TESE

SOARES, V. R. M.; ZÂNIRO, D. L.; FARIA, L. I. L. de. Análise sobre colaboração científica a partir de publicações indexadas na plataforma Dimensions. In: V Congresso Brasileiro Interdisciplinar em Ciência e Tecnologia (CoBICET), 2024. **Anais** [...]. ISSN: 2764-0582. Disponível: <https://www.even3.com.br/anais/cobicet2024/891758-analise-sobre-colaboracao-cientifica-a-partir-de-publicacoes-indexadas-na-plataforma-dimensions>. Acesso em: 04 nov. 2025.

Dedico esta tese ao avô da minha esposa, José Mariano da Silva

AGRADECIMENTOS[1]

Primeiramente, à Deus, por ter me guiado.

À minha amada esposa Silmara Corrêa da Silva Micali, pela paciência, carinho e compreensão. Seu apoio foi fundamental e fez tudo ficar mais leve.

Ao professor Dr. Leandro Innocentini Lopes de Faria, pela orientação.

Ao professor Dr. José Eduardo dos Reis, pelas valiosas dicas.

Aos colegas do NIT-Materiais/UFSCar, em especial, Dênis Leonardo Zâniro, Fernando de Natali Frascá e Denilson de Oliveira Sarvo, pela parceria e amizade.

Ao “Grupo 1” da minha turma, pela troca de conhecimentos e momentos de descontração.

Aos professores Dr. Roniberto Morato do Amaral, Dra. Luciana de Souza Gracioso, Dr. Fernando de Assis Rodrigues e Dr. Jesús Pascual Mena-Chalco, pelas contribuições no exame de qualificação e na defesa.

Muito obrigado a todos!

[1] O autor agradece à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES, pelo apoio financeiro concedido, conforme art. 1º da Portaria nº 206, de 4 de setembro de 2018.

“Feliz aquele que transfere o que sabe e aprende o que ensina”

Cora Coralina

RESUMO

A Plataforma Lattes consolidou-se como uma base de dados fundamental para a produção de indicadores de Ciência e Tecnologia no contexto brasileiro. Com o objetivo de analisar a endogamia acadêmica nas coautorias científicas dos pesquisadores da Universidade Federal de São Carlos, esta pesquisa desenvolveu um procedimento computacional integrado, capaz de articular diferentes tecnologias para extração, tratamento, processamento e visualização de dados bibliométricos. A metodologia envolveu a coleta automatizada de dados da Plataforma Lattes, o tratamento de metadados por meio de *scripts* em Python, o mapeamento das redes de coautoria científica a partir de uma aplicação em Java e a geração de visualizações com recursos de geolocalização. O conjunto de dados analisado compreende 86.341 registros de coautorias em artigos completos de revistas até o ano de 2024, correspondentes a 24.612 produções únicas, envolvendo 664 pesquisadores da UFSCar – Campus São Carlos credenciados em Programas de Pós-Graduação da universidade e 7.322 coautores com endereço profissional cadastrado no Currículo Lattes. Os resultados confirmam a hipótese da pesquisa ao mostrar que o procedimento desenvolvido explicitou as relações de colaboração que são ocultas na interface padrão da Plataforma Lattes. As evidências indicam o predomínio de coautorias não endogâmicas e a heterogeneidade de endogamia conforme a unidade de análise, além de não indicarem associações lineares entre endogamia e distância geográfica. A pesquisa mostra a eficácia do procedimento computacional para o diagnóstico empírico da endogamia e destaca o potencial das visualizações para a análise das dinâmicas de colaboração científica sob uma perspectiva interdisciplinar.

Palavras-chave: Coautoria científica. Endogamia acadêmica. Visualização de Informação. Geolocalização. Plataforma Lattes.

ABSTRACT

The Lattes Platform has established itself as a fundamental database for the production of Science and Technology indicators in the Brazilian context. Aiming to analyze academic inbreeding in the scientific co-authorships of researchers at the Federal University of São Carlos (UFSCar), this research developed an integrated computational procedure capable of coordinating different technologies for the extraction, treatment, processing, and visualization of bibliometric data. The methodology involved automated data collection from the Lattes Platform, metadata treatment through Python scripts, mapping of scientific co-authorship networks using a Java application, and the generation of visualizations with geolocation features. The analyzed dataset comprises 86,341 co-authorship records in full journal articles up to the year 2024, corresponding to 24,612 unique productions, involving 664 researchers from UFSCar – São Carlos Campus accredited in graduate programs, and 7,322 co-authors with professional addresses registered in the Lattes Curriculum. The results confirm the research hypothesis by showing that the developed procedure made explicit the collaboration relationships that are hidden in the standard interface of the Lattes Platform. Evidence indicates a predominance of non-inbred co-authorships and heterogeneity in inbreeding levels according to the unit of analysis, in addition to showing no linear associations between inbreeding and geographical distance. The research demonstrates the effectiveness of the computational procedure for the empirical diagnosis of academic inbreeding and highlights the potential of visualizations for analyzing the dynamics of scientific collaboration from an architectural and interdisciplinary perspective.

Keywords: Scientific Co-authorship. Academic Inbreeding. Information Visualization. Geolocation. Lattes Platform.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface*
BCL	Base de Currículos Lattes
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CI	Ciência da Informação
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CSV	*Comma-Separated Values*
CT&I	Ciência, Tecnologia e Inovação
CTS	Ciência, Tecnologia e Sociedade
DOI	*Digital Object Identifier*
FAPESP	Fundação de Amparo à Pesquisa do Estado de São Paulo
HTML	*Hyper Text Markup Language*
ICT	Institutos de Ciência e Tecnologia
IDE	*Integrated Development Environment*
IES	Instituição de Ensino Superior
JSON	*JavaScript Object Notation*
NoSQL	*Not Only Structured Query Language*
PL	Plataforma Lattes
PPG	Programa de Pós-Graduação
TICs	Tecnologias da Informação e Comunicação
URL	*Uniform Resource Locator*
XML	*eXtensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

A evolução da ciência está ligada ao processo de divulgação e publicação de resultados de pesquisas. Uma das maneiras utilizadas para realizar a mensuração da atividade científica é via análise de indicadores, e pela bibliometria são feitos estudos quantitativos relacionados às publicações científicas.

Segundo Dorta-González e Dorta-González (2010), a pesquisa científica influencia o desenvolvimento econômico e social, uma vez que é por meio da ciência que o homem busca conhecer o mundo e encontrar as respostas para diversas situações e fenômenos. Isso se tornou bastante visível durante a pandemia de *Coronavirus Disease 2019* (COVID-19), entre os anos 2020 e 2022, quando diversos órgãos, institutos, etc. de ciência e tecnologia foram acionados para o desenvolvimento de equipamentos médicos, medicamentos e vacinas.

Embora a cooperação acadêmica sempre tenha sido o alicerce do progresso científico, a pandemia atuou como um catalisador que elevou essa prática a patamares sem precedentes. Diante de desafios e oportunidades singulares, a comunidade científica mundial viu na coautoria de artigos — já consolidada historicamente — uma estratégia ainda mais vital para conferir agilidade à produção e à disseminação de saberes. Esse cenário intensificou drasticamente a colaboração entre pesquisadores de diferentes instituições e fronteiras, movidos pela urgência coletiva de decifrar e enfrentar o vírus (Felipe *et al*., 2022).

A coautoria, conforme definida por Smith (2019), refere-se à prática de múltiplos autores contribuírem de forma significativa para a elaboração de um artigo científico. Na pandemia, essa prática se intensificou devido à necessidade de reunir expertise multidisciplinar, envolvendo áreas como medicina, biologia, epidemiologia, ciência da computação, entre outras. Essa colaboração facilitou a troca de informações e o desenvolvimento de soluções inovadoras, além de promover maior rigor científico e validação dos resultados.

Entretanto, é importante destacar que a coautoria também apresenta desafios, como a necessidade de definir claramente as contribuições de cada autor. A atribuição correta de autoria é fundamental para evitar conflitos e assegurar o reconhecimento adequado do trabalho realizado por cada pesquisador.

Além disso, a pandemia evidenciou a importância de plataformas digitais abertas, que facilitaram o acesso às publicações e promoveram uma maior colaboração entre pesquisadores de diferentes regiões do mundo. A colaboração científica é fortalecida pela integração de dados, permitindo consolidar e certificar informações sobre pesquisadores, instituições e produções acadêmicas de forma confiável (Dias *et al*., 2023). De acordo com Oliveira (2020), a disseminação rápida de informações por meio de revistas de acesso aberto foi crucial para o avanço do conhecimento científico durante a crise sanitária.

Em suma, a coautoria de artigos científicos durante a pandemia mostrou ser uma estratégia eficaz para ampliar o conhecimento e acelerar a resposta à crise sanitária. A colaboração entre pesquisadores, aliada às normas éticas e técnicas, é fundamental para o avanço da ciência em tempos de emergência. Assim, a experiência adquirida nesse período pode contribuir para fortalecer a pesquisa colaborativa em futuras crises globais.

O estudo de Viana *et al*. (2023) mostra que a análise da produção científica, por meio de sistemas como o Sistema de Identificação e Gestão de Redes de Colaboração em Pesquisa (sisRedes), permite compreender a estrutura das colaborações acadêmicas e identificar comportamentos de interação entre pesquisadores e instituições, favorecendo a gestão estratégica da pesquisa. De acordo com os autores, a mensuração da produção científica, associada a indicadores de coautoria, contribui para avaliar o impacto e a relevância do conhecimento produzido, reforçando a importância da integração entre dados da Plataforma Lattes (PL) e do Portal da Transparência para um diagnóstico mais realista da atividade científica no Brasil.

Com a crescente transformação digital que temos presenciado, inclusive nas instituições arquivísticas (Silva; Gomes; Rodrigues, 2023), a produção e a disseminação da informação em ambientes digitais requerem a adesão de padrões para a representação da informação. Nessa conjuntura, há os metadados, que para Hillmann (2005), são compostos por atributos ou elementos que descrevem um recurso a ser recuperado por meio de uma busca.

Marques e Sayão (2023) propõem uma análise aprofundada sobre o papel estratégico dos metadados no contexto da ciência orientada por dados, uma vez que

os objetos digitais de pesquisa possuem um ciclo de vida complexo e prolongado, demandando diferentes tipos de metadados, que muitas vezes superam os próprios dados em volume e importância. Segundo os autores, esses metadados são cruciais para garantir a reprodutibilidade, o reuso, a interoperabilidade e a compreensão dos dados científicos, tanto por humanos quanto por máquinas, nos diversos contextos de pesquisa digital.

Diversas bases de dados têm sido tradicionalmente utilizadas para análise da produção científica, cada uma com características e limitações específicas. Bases internacionais, como Web of Science (WoS) e Scopus, oferecem amplo acesso à literatura científica e análises de citações, mas não fornecem informações detalhadas sobre orientações de doutorado e não possuem dados de endereço institucional, elementos centrais para esta pesquisa. Bases nacionais, como a Acácia[2], apresentam cobertura limitada para estudos que envolvam redes de colaboração. Iniciativas internacionais focadas em dados bibliométricos, como ORCID e Google Scholar, não disponibilizam informações estruturadas sobre orientações. Diante dessas limitações, a PL se destaca por ser uma base de dados nacional padronizada, contendo informações detalhadas sobre formação acadêmica, vínculo institucional, orientações e produção científica de pesquisadores brasileiros, permitindo reconstruir redes de coautoria e relações de orientação de forma única. Contudo, há uma carência de pesquisas sobre como apresentar visualmente essas informações, de modo a facilitar a análise e interpretação de indicadores, lacuna que orienta a presente pesquisa.

Ferramentas como Somos UFMG[3] e o weR_USP[4] exploram os metadados da PL para a geração de indicadores de Ciência e Tecnologia. Outra ferramenta bastante conhecida na academia para a extração de dados dos Currículos Lattes (CVs Lattes) é o scriptLattes, que é constituída por um *script* desenvolvido na linguagem de programação Python, que faz a extração de informações dos currículos formatados em HTML disponíveis publicamente no site da PL (Mena-Chalco; Cesar Jr., 2013).

---

[2] Plataforma Acácia. Disponível em: <http://plataforma-acacia.org>. Acesso em: 18 jul. 2024.
[3] Somos UFMG. Disponível em: <http://somos.ufmg.br>. Acesso em: 18 jul. 2024.
[4] weR_USP. Disponível em: <https://uspdigital.usp.br/datausp/publico/indicador/indicadores.jsp>. Acesso em: 18 jul. 2024.

De acordo com Freitas *et al*. (2001), a visualização de informação (InfoVis) pode ser entendida como o estudo das principais formas de representações gráficas para apresentação de informações, utilizando-se de técnicas como computação gráfica, interfaces homem-computador e mineração de dados, com o intuito de contribuir para a interpretação da informação por parte do usuário, visando à dedução de novos conhecimentos.

Um dos desafios é mensurar e entender aspectos de endogamia de uma instituição. Como exemplo, o estudo de Sampaio e Sanchez (2017) analisou as trajetórias de formação e atuação profissional de docentes das Faculdades de Educação da USP e da Unicamp, evidenciando um elevado grau de endogenia acadêmica: nas duas instituições, 77% dos professores obtiveram o doutorado na mesma instituição em que lecionam, o que indica baixa circulação interinstitucional e uma formação marcadamente homogênea. Apesar de existirem ferramentas relacionadas ao tema dessa pesquisa, como é o caso da Plataforma Acácia, é possível constatar que a questão de endogamia acadêmica necessita ser mais explorada.

O problema central que motivou esta pesquisa está relacionado à dificuldade de visualizar e interpretar dados contidos na PL. Embora seja uma base consolidada da produção científica brasileira, ela não oferece meios claros e intuitivos de apresentar características de coautoria científica. O desafio de mensurar e compreender a endogamia acadêmica dos pesquisadores reside na complexidade de identificar comportamentos de colaboração que frequentemente permanecem ocultos em grandes bases de dados.

Atualmente, observa-se que, mesmo com a grande disponibilidade de ferramentas de TICs, que permitem comunicação remota, muitas colaborações científicas ainda são fortemente influenciadas pela localização física. Essa constatação levanta a questão: É possível compreender a endogamia acadêmica e a proximidade geográfica na coautoria, pela coleta e processamento de dados já disponíveis na PL?

A hipótese desta pesquisa é que a utilização de métodos de tratamento e visualização de dados tornam explícitas relações de colaboração que permanecem ocultas na PL.

O objetivo geral da pesquisa foi desenvolver uma abordagem para extrair, tratar, processar e visualizar dados de produção científica — permitindo identificar endogamia acadêmica e proximidade geográfica — para aplicar nos pesquisadores da UFSCar – Campus São Carlos, credenciados em PPGs da universidade.

Esse objetivo pode ser desdobrado nos seguintes objetivos específicos:

- Investigar a endogamia no contexto da coautoria científica;
- Definir o conceito de coautoria científica endogâmica;
- Criar um procedimento computacional original de mapeamento de coautorias científicas;
- Analisar presença, intensidade e temporalidade das coautorias científicas endogâmicas dos pesquisadores da UFSCar – Campus São Carlos credenciados em PPGs da universidade, incluindo aspectos de geolocalização.

A metodologia adotada consistiu na extração de metadados sobre produção científica, orientação de doutorado e endereço institucional, seguida da construção de redes de coautoria utilizando técnicas e ferramentas de geolocalização e visualização de dados.

O grupo de pesquisa Núcleo de Informação Tecnológica em Materiais (NIT-Materiais)[5] da Universidade Federal de São Carlos (UFSCar) — que atua na pesquisa de prospecção tecnológica e inteligência competitiva, suas metodologias, ferramentas e aplicações para suporte ao desenvolvimento sustentável de empresas, arranjos empresariais e instituições públicas — há alguns anos tem realizado pesquisas sobre a PL, como por exemplo:

- Uma base referencial para o povoamento de repositórios institucionais por meio da coleta automatizada de metadados da PL (Matias, 2015);
- Uso dos dados dos CVs Lattes como recurso estratégico para a gestão dos PPGs (Maciel, 2018);
- A colaboração científica dos PPGs em CI brasileiros (Justino, 2019);

[5] NIT-Materiais/UFSCar. Disponível em: <https://www.nit.ufscar.br>. Acesso em: 25 dez. 2025.

- Visualização da colaboração científica de pesquisadores a partir de metadados da PL (Soares *et al*., 2020);
- O impacto da formação docente internacional na produção científica (Reis, 2021);
- Análise da produção de patentes no Brasil (Zâniro; Quoniam, 2025).

Mediante pesquisa realizada no NIT-Materiais/UFSCar e utilizando dados de produção científica provenientes da PL por meio do uso das ferramentas automatizadas synclattes (Matias, 2015) e csv_lattes (Matias; Amaral; Matias, 2017), foram desenvolvidos algoritmos utilizando a linguagem de programação Java, visando à obtenção de um conjunto de dados que serviram de *input*, para, com base nos conceitos da área de visualização de informação, criar visualizações gráficas em Python para análise de indicadores de endogamia acadêmica, considerando também aspectos de geolocalização.

Como motivação pessoal do autor da presente pesquisa, o tema se mostrou bastante intrigante, uma vez que, além de CTS, envolvem conceitos de Computação e CI, que são as áreas de formação do pesquisador. Além disso, as questões de originalidade e inovação que se pretendem alcançar, são aspectos fundamentais que incentivaram o desenvolvimento da pesquisa.

Esta tese está organizada em cinco seções, começando pela contextualização do problema e concluindo com as contribuições da pesquisa. A primeira seção apresenta a introdução, situando o problema, a hipótese, os objetivos e a relevância da pesquisa no contexto dos estudos métricos da ciência. A segunda seção é dedicada ao referencial teórico, explorando conceitos de coautoria científica, aspectos de endogamia acadêmica e, por fim, fundamentos de visualização de informação e *big data*, cruciais para compreender e interpretar grandes volumes de dados. Na terceira seção, descreve-se a metodologia, detalhando a extração de dados da PL, a seleção de metadados e o procedimento utilizado para explorar aspectos de endogamia e geolocalização nas coautorias científicas. A quarta seção apresenta a discussão dos resultados, por meio da interpretação das visualizações desenvolvidas. Por fim, a quinta seção traz as considerações finais, sintetizando as principais contribuições da pesquisa e apontando caminhos para pesquisas futuras.

2 REFERENCIAL TEÓRICO

Freire (2006) menciona que a criação da tecnologia de impressão facilitou a circulação da informação, e, a partir disso, surgiram os primeiros periódicos científicos e o processo de comunicação científica se constituiu, sendo que “a comunicação científica envolve a utilização de sistemas de informação para gerenciar e disseminar a informação” (Rodrigues; Rodrigues, 2023).

Segundo Bush (1945), com o aumento das publicações científicas criou-se uma dificuldade em relação ao uso eficaz do conhecimento gerado, aonde para ele, a maneira a qual o ser humano faz uso dos conhecimentos é mais importante que a consulta que é utilizada para a extração dos dados.

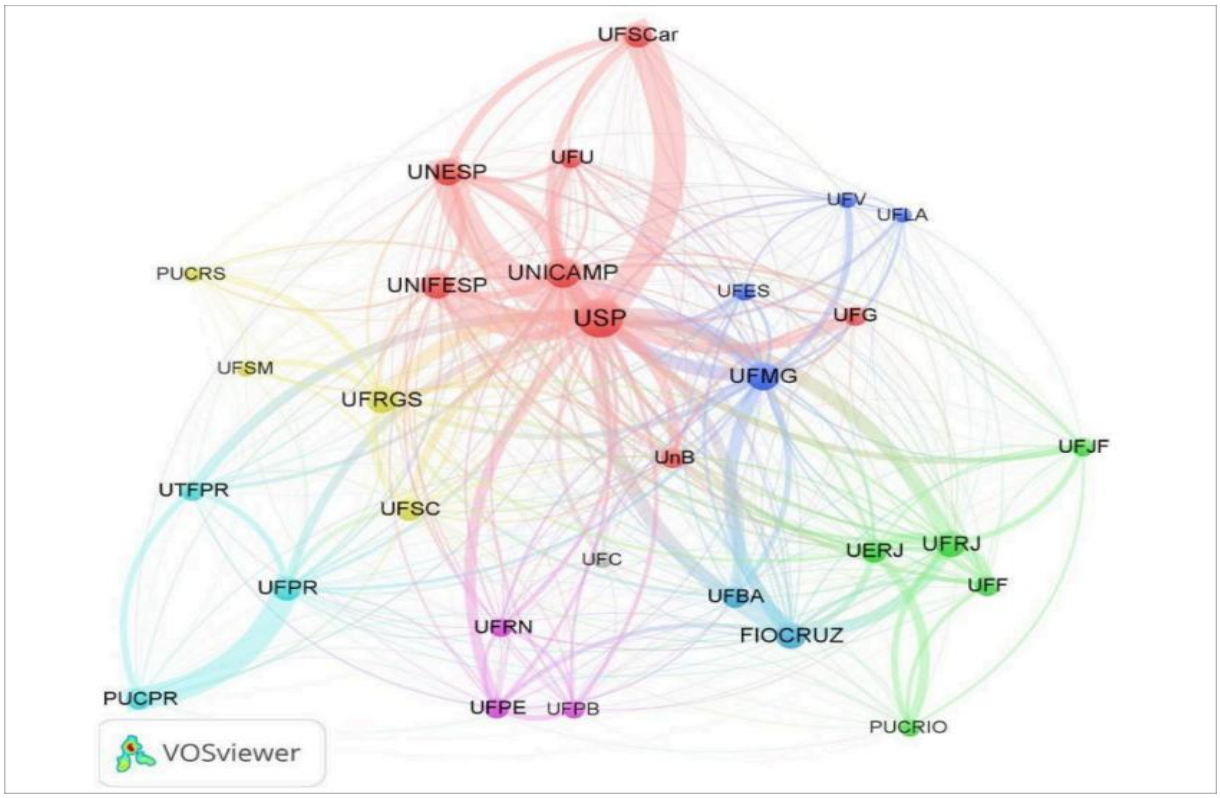
O avanço da Ciência da Computação e TICs tornou possível a criação de novas ferramentas digitais, e, por consequência, potencializou à CI no desenvolvimento de métodos e modelos de representação, transmissão, transformação e reuso da informação.

Para Oliveira *et al*. (2019), “a atual configuração da dinâmica relativa à produção e à comunicação científica permite que se revele o protagonismo da Ciência Orientada a Dados, em concepção abrangente, representada principalmente por termos como e-*Science* e *Data Science*”.

De acordo com Engwall, Blockmans e Weaire (2014), a geração e a disseminação de conhecimento por meio de ensino e pesquisa constituem-se duas obrigações básicas das instituições acadêmicas. Em relação à pesquisa, a colaboração científica é bastante importante, sendo que Boutin *et al*. (1996) argumentam que ela se torna evidente pela análise de coautoria entre pesquisadores, instituições e países.

A Figura 1 apresenta a rede de coautoria institucional construída a partir da análise de publicações de universidades e instituições de pesquisa brasileiras, visualizada com o *software* VOSviewer. Cada nó representa uma instituição, e as conexões entre elas indicam a ocorrência de coautorias em trabalhos científicos. As cores agrupam instituições com maior densidade de colaboração entre si, revelando a formação de comunidades científicas regionais e eixos de integração nacional.

Figura 1 – Rede de colaboração científica institucional no Brasil.



Fonte: Groenner *et al*. (2022).

Observa-se que o eixo paulista — formado principalmente por USP, UNICAMP, UNESP e UNIFESP — constitui o núcleo central da rede, com alta densidade de conexões e forte interligação interna. Essa configuração reflete a posição historicamente dominante dessas universidades na estrutura científica brasileira, sustentada por investimentos consistentes em pesquisa, políticas estaduais de fomento (como as da FAPESP) e infraestrutura laboratorial consolidada. A USP destaca-se como o principal nó articulador da rede, conectando-se intensamente a instituições de outras regiões e desempenhando um papel de intermediação na circulação do conhecimento científico.

Outros agrupamentos regionais também se evidenciam. O cluster fluminense, formado por UFRJ, UERJ, UFF, FIOCRUZ e PUC-Rio, apresenta forte coesão interna, representando um ecossistema de pesquisa consolidado na região Sudeste. O cluster sulista, que reúne UFRGS, UFSC, UFPR, UTFPR e PUC-RS, mostra alta densidade de conexões regionais e integração com o núcleo paulista. Já o cluster nordestino, com UFPE, UFRN, UFPB, UFBA e UFC, indica colaborações relevantes, ainda que menos centrais, refletindo esforços de integração científica em contextos

de menor densidade de infraestrutura. Por fim, observa-se a presença de nós intermediários, como a FIOCRUZ e a UnB, que exercem papel estratégico de interconexão entre diferentes regiões e áreas do conhecimento.

Sob a perspectiva dos Estudos Sociais da Ciência e da Tecnologia, essa rede expressa não apenas a distribuição das colaborações científicas, mas também as assimetrias estruturais do sistema nacional de ciência e tecnologia. A concentração das conexões em torno das universidades do Sudeste evidencia a persistência de um modelo centralizado de produção científica, em que a proximidade geográfica, a disponibilidade de recursos e a tradição institucional são fatores determinantes da cooperação. Essa tendência é consistente com análises realizadas por Sidone, Haddad e Mena-Chalco (2016), que identificaram padrões semelhantes de centralização e regionalização das colaborações científicas brasileiras a partir de dados extraídos da PL. Assim, a visualização da rede de coautoria permite compreender como a geografia da ciência brasileira reflete e reforça desigualdades históricas na formação e difusão do conhecimento.

Reis (2021) aponta que a colaboração científica acontece por diversas maneiras, como por meio do compartilhamento de equipamentos, infraestrutura e conhecimento, além da coautoria, a qual pode ser considerada em níveis: do individual aos grupos de pesquisa, departamentos, instituições, áreas e nações. Não é toda colaboração que resulta em publicação e não é toda coautoria que implica em cooperação efetiva (Gusmão; Santos; Mena-Chalco, 2022).

Sobre colaboração entre diferentes atores, observa-se o florescimento da Teoria Ator-Rede, em que fatos ou artefatos não podem ser criados de forma isolada sem redes para apoiá-los, uma vez que “o que parecem ser peças independentes de ciência e tecnologia são sempre partes de redes mais amplas” (Sismondo, 2010).

Penfield *et al*. (2014) apontam que o impacto de pesquisas pode ser influenciado por meio de redes complexas que interagem, como pesquisadores, instituições e partes interessadas externas, podendo resultar no desenvolvimento de novas ideias e produtos.

> Identificar como a disseminação do conhecimento permeia nosso meio acadêmico é uma forma de contribuir para que as relações já existentes sejam expandidas, que novas relações sejam criadas e que se pense na formação de uma rede estruturada (Braga; Gomes; Ruediger, 2008).

As redes de colaboração científica funcionam como espaços de inovação e mobilização social, podendo se configurar como redes de troca de informação, que utilizam a Internet para disseminar ideias em diferentes áreas do conhecimento, ou como redes operativas, que vão além da comunicação, atuando para desenvolver pesquisas, capacitações etc., permitindo que diversos atores unam seus esforços em torno de objetivos compartilhados (Martinho, 2003).

A colaboração científica, analisada a partir da estrutura de redes de coautoria, mostra-se essencial para compreender os processos de avaliação por pares e identificar potenciais conflitos de interesse. O estudo de Rodrigues, Tomassini e Mena-Chalco (2023) construiu redes de sugestão de pareceristas com base em vínculos de coautoria, revelando que, apesar da predominância de relações indiretas entre autores e avaliadores sugeridos, há significativa conectividade que pode comprometer a imparcialidade das avaliações.

No âmbito da medição da comunicação escrita, segundo Reis (2016), ela está associada a:

- Compreensão da dinâmica da ciência e dos fatores que definem a sua evolução;
- Planejamento, acompanhamento e avaliação de políticas públicas;
- Planejamento estratégico de empresas e instituições;
- Prestação de contas à sociedade.

Diante do cenário competitivo e da crescente complexidade da gestão universitária, as instituições de ensino superior têm buscado formas mais eficazes de monitoramento e avaliação de desempenho. Zanin (2014) propõe um painel de indicadores estruturado em quatro grandes áreas — ensino, pesquisa, extensão e gestão — que contempla 118 indicadores qualitativos e quantitativos, construídos com base em análise teórica e validação empírica junto a gestores dessas instituições. A proposta visa alinhar os processos institucionais ao planejamento estratégico de longo prazo, permitindo a melhoria contínua da qualidade educacional e a sustentabilidade organizacional, além de oferecer subsídios para enfrentar os desafios da profissionalização da gestão e da exigência por resultados mensuráveis.

Nesse contexto, é importante relatar a influência dos rankings universitários no cenário educacional global, que reflete mudanças profundas na forma como instituições de ensino superior são avaliadas, escolhidas e geridas. Righetti (2016) aponta que essas classificações, embora recentes, exercem impacto direto nas decisões de alunos, nas estratégias de marketing institucional e até mesmo em políticas públicas. Os rankings, frequentemente produzidos por grupos de mídia, legitimam um modelo em que instituições com maior produtividade científica tendem a obter mais recursos, reforçando desigualdades e consolidando uma lógica meritocrática baseada em critérios muitas vezes questionáveis.

A crescente valorização dos rankings universitários reflete o papel estratégico que a produção científica desempenha na reputação internacional das universidades. Segundo Santos (2015), o desempenho das universidades brasileiras nesses rankings está fortemente associado à sua produção científica. A autora observa que, embora a participação brasileira na produção científica global tenha aumentado significativamente, esse crescimento não se traduz diretamente em melhores posições nos rankings internacionais, evidenciando que outros critérios, além da produção acadêmica, influenciam essas classificações.

Para Van Raan (2019), o estudo quantitativo da ciência, geralmente referido como cienciometria, visa o desenvolvimento da ciência e sua comunicação, em relação a aspectos tecnológicos, sociais e socioeconômicos, utilizando ligações interdisciplinares com a Filosofia, História, Sociologia da Ciência, Política, Administração, Matemática, Física e CI.

Amaral, Matias e Sarvo (2024) relatam a presença de interdisciplinaridade na CI ao identificar que 74% dos docentes dos PPGCI brasileiros publicam em coautoria com profissionais de outros campos, chegando a artigos que envolvem até 14 áreas distintas. Além disso, eles descobriram que a intensidade dessas interações se revela no núcleo interdisciplinar formado por Ciência da Computação, Educação, Administração, Comunicação e Museologia, áreas presentes em aproximadamente 80% das publicações analisadas. Já a temporalidade dessa interdisciplinaridade evidencia sua evolução progressiva entre 2013 e 2020, período marcado pelo crescimento contínuo das coautorias e pela diversificação das áreas de atuação autodeclaradas na PL, revelando que a articulação entre saberes acompanha transformações históricas, tecnológicas e institucionais do campo,

configurando a interdisciplinaridade como fenômeno dinâmico e estruturante da identidade epistemológica da CI no Brasil (Amaral; Matias; Sarvo, 2024).

Segundo Wyatt *et al*. (2017), a cientometria e as abordagens qualitativas dentro do CTS possuem uma origem comum, mesmo levando em consideração o fato delas terem se distanciado nas últimas décadas em termos de práticas, normas e padrões de pesquisa, até porque, diferentes habilidades são necessárias aos pesquisadores e os pressupostos epistemológicos também são diferentes.

> Considerando os recursos humanos científicos atuantes em pesquisas cientométricas, um aspecto importante que caracteriza a maioria desses grupos é a composição multidisciplinar de seus membros [...] sendo necessário montar grupos com pessoas de diversas especialidades. (Costas, 2017).

“Existem diversas formas de medição voltadas para avaliar a ciência e os fluxos da informação” (Vanti, 2002). Entre elas, a técnica da bibliometria pode ser utilizada para a realização de estudos quantitativos relacionados às publicações científicas, conhecidos por indicadores científicos. A bibliometria surge como ferramenta essencial para o acompanhamento da produção científica, permitindo identificar tendências emergentes, características de colaboração e dinâmicas de publicação, possibilitando, entre outras coisas, a classificação de temas, análise de redes de coautoria e visualização de padrões de evolução científica, o que a torna estratégica para o processo de prospecção tecnológica (John; Fritsche, 2013).

O avanço das tecnologias computacionais possibilitou uma maior disponibilidade de bases de dados eletrônicas de periódicos e facilitou a extração, o armazenamento e o tratamento de dados, como também, tornou os artigos científicos o tipo de fonte mais utilizado a nível global na construção de indicadores (Faria et al., 2011). Costas (2017) aponta que o aumento da capacidade de computação e de armazenamento, juntamente com a disponibilidade de dados bibliográficos, fez com que crescesse o número de pessoas com acesso a indicadores, ferramentas e aplicações, baseados na bibliometria. É importante ressaltar que “os dados desejados podem estar disponíveis apenas através de fornecedores que exigem assinaturas [...] e os preços podem ser proibitivos” (Wolfram, 2017).

A produção científica é compreendida como um dos principais indicadores do desenvolvimento tecnológico, uma vez que novas tecnologias podem se apoiar em

descobertas e avanços realizados em laboratórios de pesquisa. Okubo (1997) argumenta que os indicadores de produção científica são indispensáveis para se discutir a respeito dos avanços da ciência e da tecnologia. No Brasil, a produção de ciência praticamente em sua totalidade é realizada dentro das universidades, ou seja, o conhecimento gerado por essas instituições é de grande importância para o avanço econômico e científico do país (Santos, 2015).

De acordo com Spinak (1998), a avaliação da ciência é uma parte da política pública de um país, pois ela objetiva medir o desempenho científico e verificar se o esforço despendido em pesquisas, que resulta em publicações, realmente contribuiu para o progresso da sociedade, usando como referência as metas de política científica e tecnológica que foram determinadas para o país ou região.

Segundo Bassoli (2017), a produção do conhecimento e da ciência está diretamente relacionada ao potencial de divulgação de resultados, sendo que a validação das pesquisas, dos pesquisadores e das instituições acontece regularmente por meio de mecanismos de avaliação da qualidade, do volume, da relevância e do alcance das publicações.

Targino (2000) aponta que “a comunicação científica é indispensável à ciência, pois permite somar os esforços individuais dos membros das comunidades científicas”. Lima (2007, p. 29) argumenta que “a publicação científica tem sido, historicamente, a fonte de dados mais utilizada para gerar indicadores que permitam analisar os resultados e a qualidade da produção científica e, ainda, estimar o impacto científico”.

De acordo com Penfield *et al*. (2014), as métricas têm sido comumente utilizadas como uma medida de impacto também fora do âmbito acadêmico, por exemplo, em relação a lucro obtido, número de empregos e de contratações, quantidade de vendas etc., sendo frequentemente vistas como evidências inequívocas. No caso da produção científica, antes centrada apenas em publicações e citações, ganhou maior amplitude com o banco de dados Dimensions, que integra financiamentos, ensaios clínicos, patentes, documentos de políticas públicas e métricas alternativas, desse modo, ampliou-se a compreensão dos impactos da ciência ao favorecer análises em larga escala, contribuindo para maior inovação e transparência nos estudos métricos (Herzog; Hook; Konkiel, 2020).

Os indicadores “podem ser compreendidos como dados estatísticos aproveitados, para medir algo intangível, que ilustram aspectos de uma realidade multifacetada”, como explica Gregolin *et al*. (2005). De acordo com Prado e Castanha (2020), os indicadores são recursos estratégicos para serem utilizados em diversos aspectos que necessitem compreender conjunturas e cenários determinados. Para isso, os indicadores precisam ser analisados, de modo a transformar dados em informação.

Para Glänzel, Thijs e Debackere (2019), os métodos quantitativos em estudos científicos — como estatísticas, medidas e indicadores — capturam e expressam aspectos importantes e característicos sobre uma quantidade relevante de objetos, geralmente empregando diversas metodologias e técnicas.

Para Davyt e Velho (2000), os principais indicadores, que são derivados das publicações de artigos científicos, são divididos em três categorias:

- Indicadores de publicação: medem a quantidade de pesquisas publicadas e o impacto das publicações;
- Indicadores de citação: medem a qualidade e o alcance das publicações por meio das citações por outras pesquisas;
- Indicadores de ligação: medem as relações entre os coautores das pesquisas, seus grupos de pesquisa e instituições.

Para Silva et al. (2023), em termos gerais, se um artigo foi citado, infere-se que ele foi lido e considerado pertinente para o embasamento de novas pesquisas. Portanto, entende-se que os estudos sobre comunicação científica e avaliação da ciência possuem grande importância para a compreensão dos impactos das políticas públicas envolvendo Ciência, Tecnologia e Sociedade.

A respeito de indicadores de ligação, Narin, Olivastro e Stevens (1994) argumentam que eles mensuram as coocorrências de autores, afiliação, palavras e referências. Nessa conjuntura, “a análise baseada em ligações com base em citações (citação direta, cocitação, acoplamento bibliográfico) ou coautoria tem sido um ponto focal para a pesquisa métrica durante décadas” (Wolfram, 2017). A análise de citação permite compreender como ideias científicas evoluem e se disseminam ao longo do tempo, oferecendo uma visão estrutural da rede de conhecimento

(Small, 2004). Técnicas de cocitação são úteis para identificar agrupamentos temáticos e núcleos de pesquisa com afinidade conceitual (Small, 2004). Por sua vez, o acoplamento bibliográfico, mensura a relação entre dois artigos considerando o número de referências em comum citadas por ambos (Kessler, 1963).

Há também métricas alternativas, conhecidas como altmetria, que segundo Puerta-Díaz, Martí-Lahera e Martínez-Ávilla (2020), “compreende um conjunto de métricas para mensurar o impacto das publicações de maneira complementar à bibliometria tradicional”. A altmetria consiste no uso de métricas alternativas que permitem acompanhar como os resultados de pesquisas são acessados e utilizados em ambientes digitais, oferecendo uma visão complementar à avaliação tradicional baseada apenas em citações (Nascimento, 2016).

Além de ampliar a compreensão sobre a circulação e relevância dos trabalhos científicos, a altmetria também aponta tendências e favorece uma avaliação mais abrangente da produção científica (Araujo, 2018). A complementaridade entre indicadores bibliométricos e altmétricos fornece uma visão mais abrangente do impacto da produção científica, tanto no meio acadêmico quanto na sociedade (Gontijo; Araújo, 2021).

A partir de um estudo realizado por Lyu e Costas (2020), a colaboração científica foi evidenciada ao mostrar como os temas acadêmicos se deslocam entre a comunidade científica e os diferentes públicos online, indicando que encontros síncronos virtuais “não apenas reinterpretam os tópicos acadêmicos, mas também os ampliam, relacionando-os a questões sociais e práticas mais amplas” (Lyu; Costas, 2020, p. 928).

As métricas alternativas podem medir o impacto das publicações científicas levando em consideração o engajamento de usuários. Elas foram viabilizadas pela Web 2.0, e, principalmente, com o advento das redes sociais. “Uma rede social é um grupo de pessoas, de organizações ou de outros relacionamentos, conectados por um conjunto de relações sociais, como as amizades, o trabalho em conjunto ou a simples troca de informações” (Braga; Gomes; Ruediger, 2008). Nesse contexto, as métricas alternativas indicam mais o alcance social e o potencial de visibilidade do que, necessariamente, a qualidade científica de uma pesquisa (Boon; Foon, 2014).

Para Rodrigues e Rodrigues (2023), a não compreensão das perspectivas a respeito de pesquisas sobre redes sociais online constitui um fator limitante para um melhor entendimento sobre a exploração dessas plataformas pela academia, ou seja, elas necessitam serem consideradas nos estudos métricos. Freitas, Benchimol e Rodrigues (2023) analisaram publicações no Twitter a respeito do evento MuseumWeek[6] e evidenciaram que os serviços de redes sociais online se constituem como espaços de interações interinstitucionais, engajamento público e compartilhamento de conteúdos científicos e culturais em escala global.

Gouveia e Araújo (2020) apontam que, ao se fazer uma postagem em uma rede social, é gerado um elemento que possui vínculos com o autor e conexões com as pessoas que interagem com ela por meio de curtida, compartilhamento etc. Rodrigues (2024) destaca que serviços de redes sociais online funcionam como sistemas complexos, estruturados para coletar, armazenar e disponibilizar dados, possibilitando tanto o desenvolvimento de estudos quanto a cooperação entre pesquisadores de diferentes áreas e localidades.

Pensando na questão de recrutamento de pesquisadores em IES, Borenstein, Perlin e Imasato (2022) argumentam que, embora o processo de seleção seja formalmente baseado no mérito e aberto à concorrência, há percepções de que os laços informais podem influenciar as decisões de contratação, ou seja, os autores mencionam a importância das redes sociais. Para Krackhardt (1990), a estrutura social, que inclui a forma como as pessoas estão organizadas e conectadas, é uma base fundamental para a ação e a influência.

O estudo de Godechot e Louvet (2008) mostra que candidatos endogâmicos apresentam probabilidade significativamente maior de serem recrutados em relação aos externos. Os autores argumentam que tal característica tende a limitar a mobilidade e a diversidade acadêmica, podendo resultar na rejeição de candidatos mais qualificados e, consequentemente, comprometer a qualidade geral dos processos de recrutamento nas universidades.

Segundo Borenstein, Perlin e Imasato (2022), existe uma variável denominada como herança acadêmica que representa o conhecimento intangível,

---

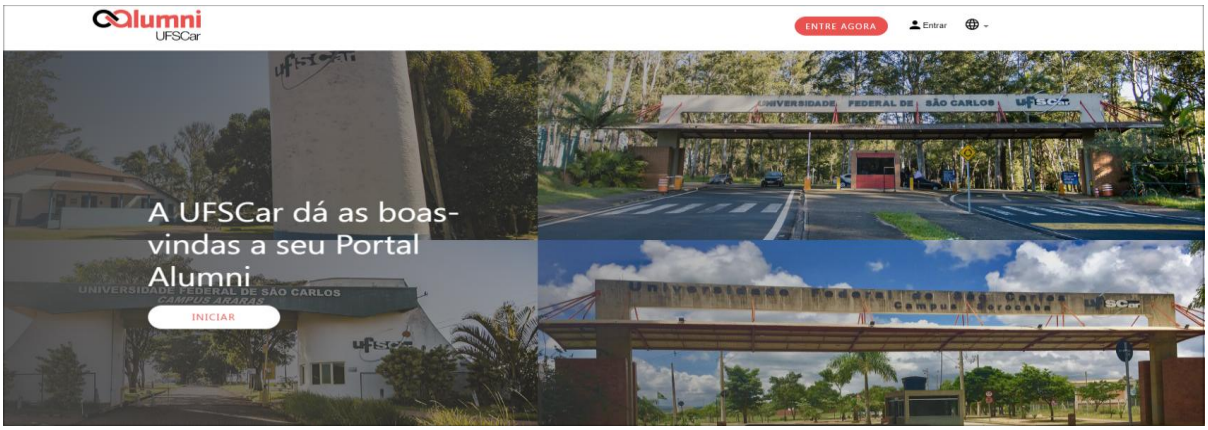
[6] MuseumWeek. Disponível em: <https://en.wikipedia.org/wiki/Museum_Week>. Acesso em: 22 ago. 2025.

considerando, entre outros fatores, o *networking* acadêmico de um pesquisador durante seu período de formação em uma instituição. Para Rocca (2007), há contratação de profissionais em que se priorizam as conexões sociais mais do que os méritos acadêmicos.

Recentemente foi lançado o Portal Alumni da UFSCar[7], que possibilita a reconexão entre colegas de turma, visando expandir as relações profissionais (*networking*), para, entre outras possibilidades, ofertar mentoria a estudantes. No futuro, é possível vislumbrar que essa plataforma possa ser uma rica fonte de dados para geração de indicadores.

A Figura 2 mostra a página inicial do Portal Alumni da UFSCar.

Figura 2 – Portal Alumni da UFSCar.



Fonte: Portal Alumni da UFSCar, acesso realizado pelo autor.

No âmbito de dados disponíveis em plataformas digitais, Matias (2015) destaca que além dos indicadores existentes em bases de dados reconhecidas pela comunidade científica, como a Scopus e a WoS, há a possibilidade de extração de indicadores de outras bases. “Além das bases de dados internacionais, a maioria dos grupos de pesquisa também utiliza fontes de dados de interesse local ou temático” (Costas, 2017).

A PL se destaca como um sistema de informação na Web sob responsabilidade do CNPq, que permite a integração das bases de dados de Currículos, Grupos de Pesquisa e Instituições, agregando currículos de estudantes e pesquisadores do país (Autran *et al*., 2015).

[7] Portal Alumni da UFSCar. Disponível em: <https://alumni.ufscar.br>. Acesso em: 18 jul. 2024.

A BCL é utilizada por universidades, institutos, centros de pesquisa e fundações de amparo à pesquisa dos estados como mecanismo para a avaliação de pesquisadores, professores e alunos, estabelecendo-se como uma grande base de dados de pesquisadores do país. Também é estratégica para a formulação das políticas do Ministério da Ciência, Tecnologia e Inovações (MCTI) e de outros órgãos governamentais da área de Ciência, Tecnologia e Inovação (CNPq, 2023).

De acordo com uma auditoria realizada pelo Tribunal de Contas da União, em setembro de 2022 a PL tinha 7,7 milhões de currículos, 30 mil grupos de pesquisa certificados e 40 mil instituições cadastradas (TCU, 2023).

Os CVs Lattes mantêm referências a documentos públicos e privados dos cientistas, sendo um histórico das atividades científicas, acadêmicas e profissionais de pesquisadores cadastrados, as quais são livremente inseridas, de forma a representar a produção científica em vários meios, como: artigos em revistas internacionais, com fator de impacto e indexadas em bases de dados; artigos em revistas locais, sem DOI, de baixa circulação; livros, capítulos de livros, trabalhos completos, resumos expandidos e resumos em anais de congressos.

A disponibilização pública dos dados da PL na Internet possibilita transparência e confiabilidade às atividades de fomento do CNPq, como também para as agências que usam a plataforma. Além disso, promovem o intercâmbio entre pesquisadores e instituições, constituindo-se como uma fonte de informações para estudos e pesquisas. Como as informações da BCL são recorrentes e cumulativas, a memória da atividade de pesquisa no país é preservada (CNPq, 2023).

Por meio de um convênio, o CNPq possibilita que as IES obtenham cópias dos CVs Lattes de seus pesquisadores (alunos e professores). Esse convênio permite a implementação de tecnologias como a ferramenta synclattes, que é um conjunto de *scripts* desenvolvidos por Matias (2015) com a linguagem de programação Python, que realiza a extração de dados dos CVs Lattes.

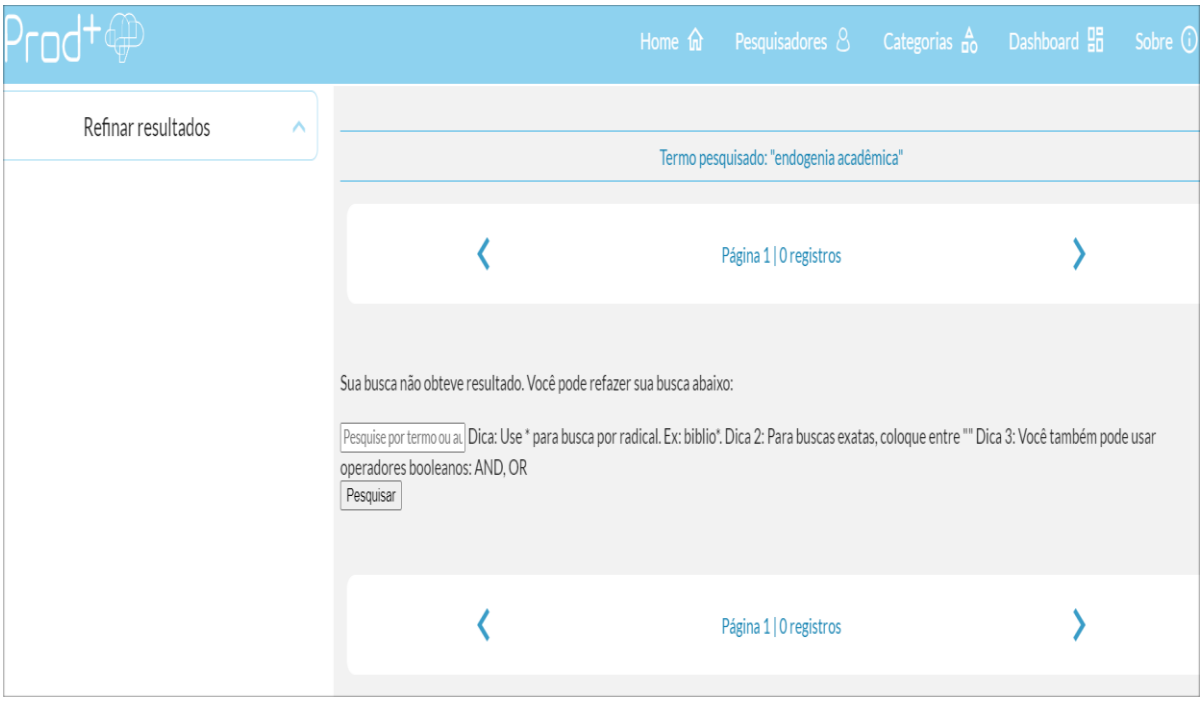
Segundo Bassoli (2017), a coleta automatizada de volumosas quantidades de metadados das publicações científicas foi essencial a potencialização das análises bibliométricas. Com a grande quantidade de informação que está disponível, necessita-se de estudos a respeito de mostrá-la por meio de indicadores

quantitativos, possibilitando representá-la num formato visual que facilite a sua compreensão.

Como exemplo de aplicação que trabalha com indicadores quantitativos de forma visual, o Prodmais[8] é um *software* livre cujo código-fonte está disponível no GitHub[9], que foi desenvolvido originalmente pela Universidade Federal de São Paulo para universidades e centros de pesquisa. É uma ferramenta que agrega informações sobre produções acadêmicas de várias fontes, principalmente da PL, que possibilita realizar buscas específicas na base de dados e filtrar os resultados, por exemplo, por área de atuação, campus, idioma, data da publicação, nível de formação etc.

A Figura 3 apresenta uma tela de busca do Prodmais, onde, a nível de ilustração, foi feita uma pesquisa por um determinado termo, de modo a mostrar o tipo de retorno que a ferramenta proporciona.

Figura 3 – Prodmais.



Fonte: Prodmais, busca realizada pelo autor.

[8] Prodmais. Disponível em: <https://unifesp.br/prodmais/index.php>. Acesso em: 18 jul. 2024.
[9] Código-fonte do Prodmais no GitHub. Disponível em: <https://github.com/unifesp/prodmais>. Acesso em: 18 jul. 2024.

Curiosamente, ao pesquisar “endogenia acadêmica”, nenhum resultado foi retornado. Dentre as hipóteses, há de se considerar que o mecanismo de busca não esteja totalmente funcional (o conjunto de dados ainda não está indexado de forma adequada), bem como pode estar ocorrendo alguma particularidade de tradução etc. Gracioso (2017) aponta que o desafio nas buscas por informação está relacionado a utilização de termos, conceitos e palavras, e a CI deve prover formas para estimular os usuários sobre as possíveis aplicações dos conteúdos buscados, visando à melhoria da vida individual e em sociedade.

Endogenia acadêmica[10] é um termo derivado da Biologia que pode ser definido como “a nomeação de docentes que se formaram na mesma instituição que os emprega” (Altbach; Yudkevich; Rumbley, 2015). Há outros termos relacionados, como “eugenia”[11], que refere-se a um conjunto de práticas e ideias com o objetivo de “melhorar” a qualidade genética da população humana.

Foi percebido que no Brasil é mais utilizado o termo “endogenia acadêmica”, em Portugal usa-se “endogamia académica” e em inglês são utilizados os termos “*academic inbreeding*” e “*intellectual inbreeding*”. Interessante ressaltar que “*inbreeding*” é traduzido para endogamia e “*breeding*” tem o sentido de cruzamento, reprodução e procriação. Sendo assim, há de se tomar cuidado na pesquisa por referências, uma vez que endogamia pode estar relacionado a parentes que procriam, por exemplo, entre animais irmãos ou primos. Já endogenia é a característica de endógeno, o novo contratado que se formou na própria instituição, ou seja, a “mãe” que contrata o seu “filho”. Nesse contexto, optamos nesta pesquisa por utilizar o termo endogamia, no sentido de “artigo que nasce por meio de uma coautoria”.

Borenstein, Perlin e Imasato (2022) argumentam que a ideia de endogenia como uma metáfora biológica auxilia a identificar ameaças que podem ser causadas a um ambiente institucional acadêmico, porém, ela não deve ser vinculada a esse sentido biológico, pois senão poderá gerar uma interpretação equivocada, como se apenas houvessem aspectos negativos.

---

[10] Endogenia acadêmica. Disponível em: <https://en.wikipedia.org/wiki/Intellectual_inbreeding>. Acesso em: 18 jul. 2024.
[11] Eugenia. Disponível em: <https://pt.wikipedia.org/wiki/Eugenia>. Acesso em: 18 jul. 2024.

Targino (2010) aponta que os membros da comunidade científica costumam colaborar profissionalmente com instituições distintas, sendo que no Brasil isso ocorre principalmente em universidades, institutos, sociedades científicas e associações. Por sua vez,

> O papel das universidades traduz-se em efetivo compromisso com a solução de problemas e desafios de seu contexto econômico-social, implicando responsabilidades quanto a interesses e necessidades sociais. Afigura-se, portanto, o quanto as universidades são essenciais para o desenvolvimento de um país, interagindo logicamente com o poder público, o setor produtivo e a sociedade como um todo (Santos, 2015, p. 26).

A coautoria científica ocorre em múltiplas situações, como nas comunicações orais, em exposições dos resultados de estudos e na escrita de artigos, de modo a oferecer outros benefícios além da questão financeira, como crédito e reconhecimento pelos pares.

Segundo Gazda e Quandt (2010), no ambiente acadêmico a cooperação formal ou informal é um dos pilares do desenvolvimento científico, onde os vínculos são criados por meio de projetos interinstitucionais, de grupos de pesquisa, do envolvimento em PPGs e em cursos de outras instituições, da participação em simpósios, seminários e congressos, e da participação em bancas de avaliação de trabalhos. Rodrigues e Mena-Chalco (2024) realizaram um estudo que mostrou que em 67 de 83 defesas de doutorado analisadas, ao menos um membro da banca já possuía coautoria com o orientador, evidenciando como essas redes se estendem além da participação na tese.

Para Vanz (2009), colaboração e coautoria não são sinônimos, porque a coautoria é uma parte da colaboração científica, pois ela não mede a colaboração em sua complexidade e completude. Hilário e Freitas (2020) argumentam que para analisar a colaboração científica de maneira adequada deve-se realizar análises de documentos institucionais, de acesso público ou privado.

Franco e Faria (2019) apontam que pode haver colaborações informais que não são possíveis de serem verificadas apenas por meio da análise de rede baseada em publicação de artigos com coautoria, porque as relações informais não são passíveis de análise quantitativa. “Quando se fala de cooperação científica na visão da bibliometria, devemos ter em mente que quase sempre estamos nos

referindo à análise de trabalhos publicados em coautoria" (Lima; Velho; Faria, 2007, p. 54).

De acordo com Camargo Jr e Coeli (2012), em todo o mundo vem acontecendo um aumento no número de autores por artigo, possivelmente devido às pressões para que os pesquisadores publiquem cada vez mais e em decorrência da complexidade de certos tipos de estudos.

Segundo Lopes e Costa (2012), ainda que a coautoria seja tendência, existem divergências teóricas que a impossibilitam. Porém, para Gazda e Quandt (2010), a capacidade de criar novos conhecimentos muitas vezes ocorre devido a exploração da diversidade de competências complementares internas e externas à organização.

Nesse âmbito de coautoria, há de se pensar na questão de produtividade acadêmica. Para Bourdieu (1983), o pesquisador depende de sua reputação para conseguir, entre outras coisas, fundos para pesquisas, bons estudantes, convites, consultorias, prêmios, as quais se tornam mais plausíveis de serem obtidas a partir da vantagem cumulativa, que

> refere-se aos processos sociais por meio dos quais vários tipos de oportunidades de pesquisa científica, assim como as recompensas simbólicas subsequentes aos resultados daquela pesquisa, tendem a acumular-se para os praticantes individuais da ciência. [...] O conceito de vantagem cumulativa dirige nossa atenção para as maneiras pelas quais as vantagens comparativas iniciais, relativas à capacidade adquirida, localização estrutural e recursos disponíveis, contribuem para incrementos sucessivos da vantagem, de modo que as distâncias entre os que têm e os que não têm na ciência (assim como em outros domínios da vida social) ampliam-se até que sejam refreadas por processos compensatórios (Merton, 2013, p. 200).

Bourdieu considera o mundo do laboratório como uma estrutura microssociológica que conduz à noção de campo: o campo científico (Bourdieu, 2004). Para tanto, o campo "é um 'sistema' ou 'espaço' estruturado de posições ocupadas pelos diferentes agentes", como aponta Bernard Lahire (Catani *et al*., 2017). Nesse ambiente, ele é tanto um campo de forças pois pode constranger os agentes, como um campo de lutas já que os agentes atuam em suas posições conservando ou transformando a estrutura do próprio campo. Podem existir vários campos, e em cada um ocorre um determinado tipo de capital.

Quanto ao capital científico, trata-se de "um capital fundado no conhecimento e no reconhecimento. Poder que funciona como forma de crédito, pressupõe a

confiança ou a crença dos que o suportam porque estão dispostos [...] a atribuir crédito” (Bourdieu, 2004).

Em relação ao capital social, Cross, Parker e Sasson (2003) argumentam que ele é definido por sua função, surgindo por meio de mudanças nas relações entre pessoas, e, assim como outros capitais, é produtivo, possibilitando a realização de certos fins que em sua ausência não seriam possíveis.

A cooperação favorece o processo inovador, uma vez que possui a capacidade de unir as atividades práticas que acontecem nas empresas com os aspectos científicos de pesquisa e desenvolvimento que ocorrem na academia (Gazda; Quandt, 2010). Nessa conjuntura, o conceito de redes e grafos é muitas vezes utilizado tanto no âmbito empresarial quanto no acadêmico (por exemplo, para representar a genealogia acadêmica). Um grafo é definido como um conjunto de pontos e um conjunto de linhas que conectam esses pontos, as quais representam relações, como interação ou comunicação, entre os membros de uma organização (Krackhardt, 1994).

A genealogia acadêmica é definida como o estudo da herança intelectual que resulta das relações entre orientadores e orientados nos níveis de mestrado e doutorado, permitindo compreender as linhagens acadêmicas e redes de parentesco científico. Representada graficamente por árvores genealógicas ou genogramas acadêmicos, a genealogia acadêmica categoriza indivíduos de acordo com sua descendência intelectual, possibilitando o entendimento da propagação do conhecimento científico e da influência de pesquisadores na formação de novas gerações (Rossi; Mena-Chalco, 2014). Complementando,

> analisar os relacionamentos de orientação, sob a forma de uma estrutura genealógica (*e.g.*, grafo ou árvore), permite um maior entendimento sobre a comunidade científica, a caracterização do acadêmico por meio de seus relacionamentos e a identificação do impacto gerado por esses atores na constituição de seus respectivos grupos (Rossi; Damaceno; Mena-Chalco, 2018, p. 198).

Marques *et al*. (2024) realizaram uma análise da genealogia acadêmica do Prof. César Lattes, que evidenciou uma rede de descendência caracterizada pela baixa reprodução de orientadores, onde, dos 851 descendentes identificados, apenas 109 (12,8%) tornaram-se orientadores.

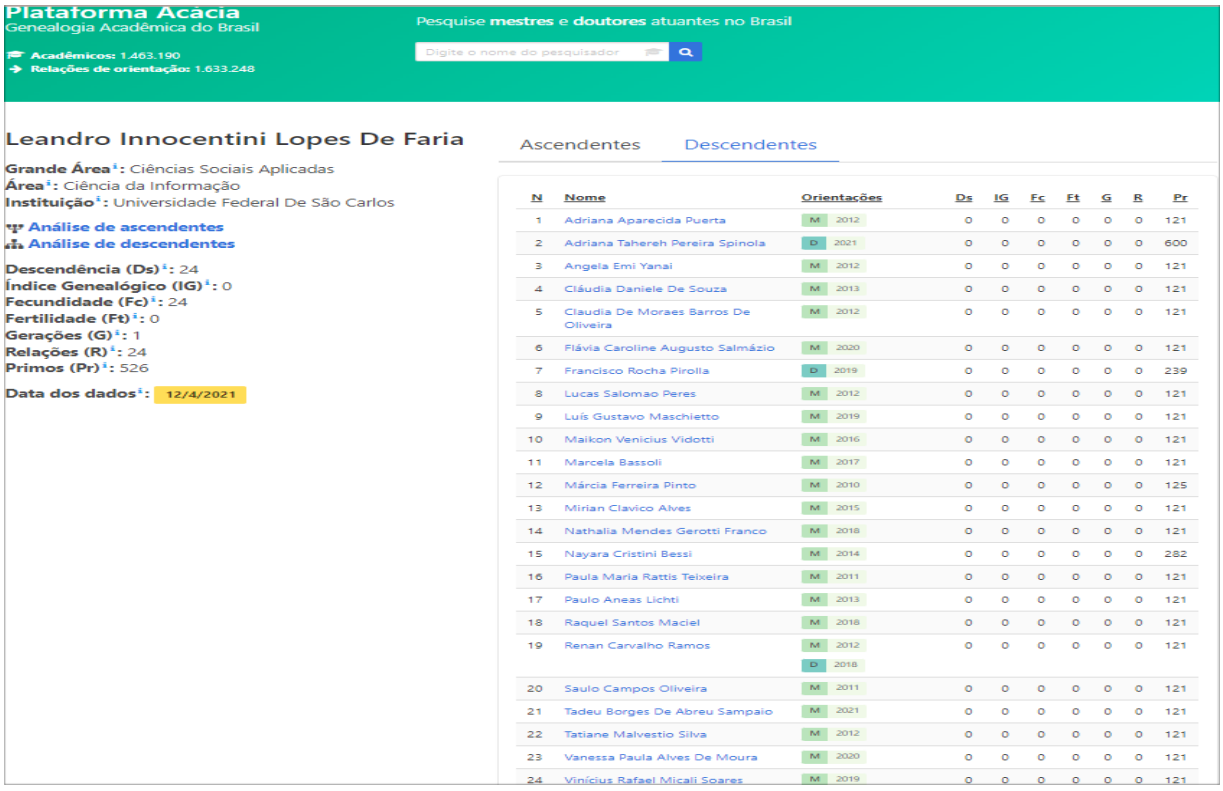
De acordo Soares, Souza e Moura (2010), a visualização de redes é composta por ligações, de modo que é possível identificar as cooperações, sendo “uma ponte crucial entre duas árvores formadas densamente por amigos próximos” (Braga; Gomes; Ruediger, 2008).

Nesse contexto de redes, podem ser utilizadas técnicas de visualização de informação, que, em suma, são responsáveis pela representação de dados de forma gráfica, a fim de que os indicadores provenientes sejam mais facilmente interpretados pelo cérebro humano.

A Plataforma Acácia, que foi criada com o objetivo de documentar as relações formais de orientação dos PPGs brasileiros, utiliza dados provenientes da PL para representar a genealogia acadêmica por meio de grafos, onde cada vértice representa um pesquisador e cada aresta uma relação de orientação concluída entre dois pesquisadores (orientador e orientado).

Na Figura 4 é apresentado o retorno da busca da Plataforma Acácia ao digitar o nome de um pesquisador.

Figura 4 – Plataforma Acácia.



Fonte: Plataforma Acácia, busca realizada pelo autor.

O processo de internacionalização da ciência tem se intensificado, sendo mensurado de forma eficiente por meio de indicadores bibliométricos. A cooperação internacional é apontada como um dos principais indicadores de desempenho da produção científica nas IFES, por refletir tanto a qualidade quanto o alcance global das publicações (Santini, 2017). Ferreira, Mcmanus e Faria (2022) argumentam que na bibliometria avaliativa foram desenvolvidos vários estudos a respeito de pesquisas colaborativas internacionais, as quais têm sido estimuladas pelas instituições de CT&I.

Reis *et al*. (2021) verificaram que a coautoria internacional é positivamente influenciada pela formação de doutorado ou pós-doutorado no exterior. A mobilidade internacional na pós-graduação, especialmente por meio do doutorado sanduíche, impulsiona o desenvolvimento acadêmico e profissional dos discentes brasileiros ao ampliar redes de pesquisa, fortalecer colaborações interinstitucionais e possibilitar o acesso a tecnologias e metodologias avançadas (Guerra; Augusto; Leão, 2025). Além disso, a mobilidade potencializa a produção científica ao ampliar o capital social dos pesquisadores, promovendo maior produtividade, impacto por citações e integração em redes de colaboração (Momeni *et al*., 2022). Em contrapartida, a endogamia acadêmica tende a reduzir a mobilidade e a abertura internacional (Altbach; Yudkevich; Rumbley, 2015), limitando a inovação das universidades (Horta; Meoli; Santos, 2021) e o reconhecimento dos profissionais (Sivak; Yudkevich, 2012).

Tavares *et al*. (2022) argumentam que a endogamia acadêmica está associada a redes de pesquisa menores e mais fechadas, concentradas em colaborações dentro da própria instituição, o que limita a diversidade e a inovação científica, ao passo que acadêmicos não endogâmicos, com doutorados no exterior, apresentam redes internacionais mais amplas. O estudo de Perlin *et al*. (2017) demonstra que pesquisadores com doutorado obtido em instituições brasileiras — ou seja, formados no próprio país — publicam mais artigos, porém em periódicos de menor impacto, enquanto aqueles com doutorado no exterior têm menor produtividade em quantidade, mas maior visibilidade internacional. Ainda segundo os autores, essa relação reflete como a formação doméstica tende a reforçar redes internas de pesquisa, limitando a inserção internacional e, portanto, favorecendo dinâmicas endogâmicas dentro da academia brasileira.

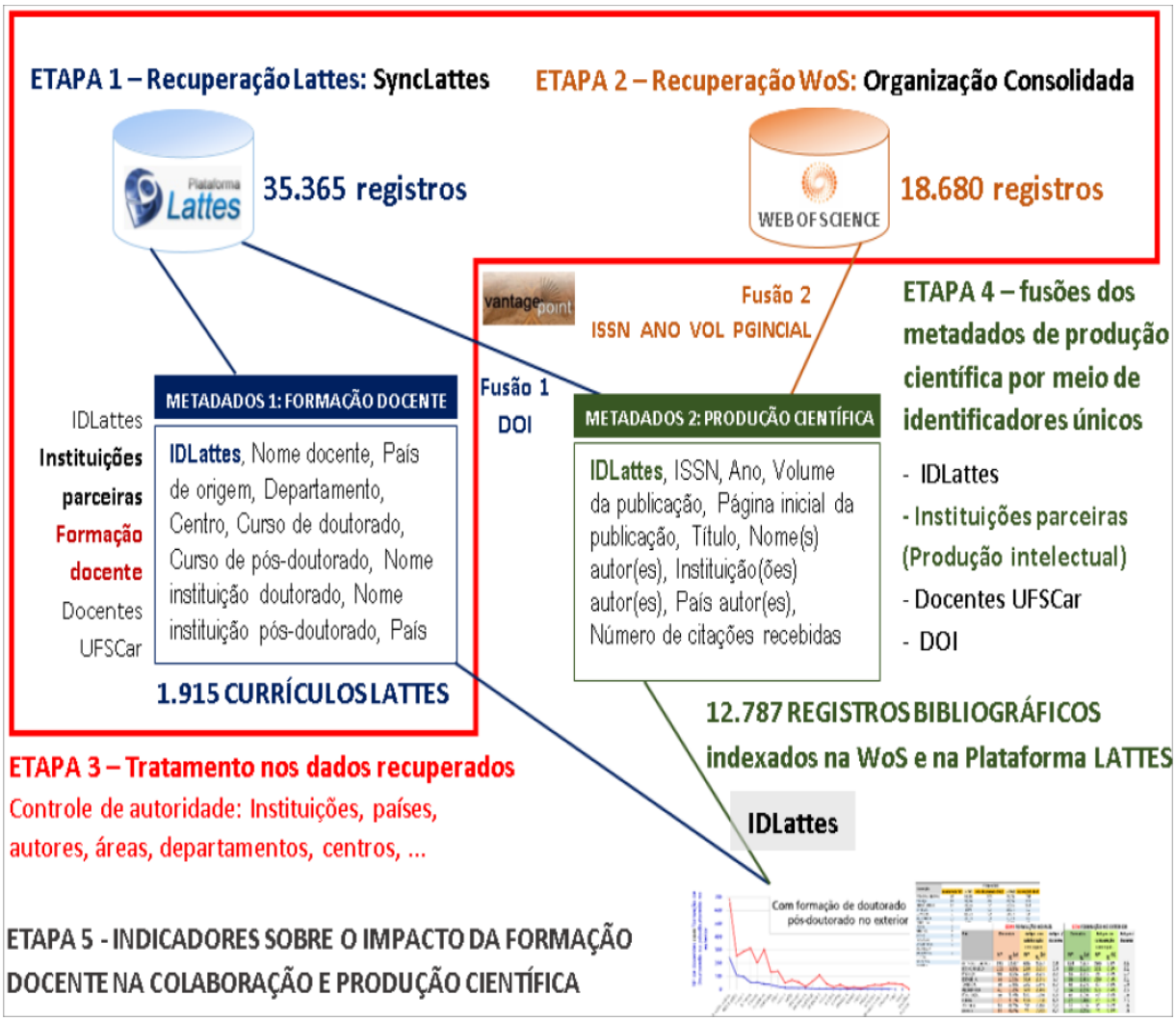
Gusmão e Mena-Chalco (2024) realizaram um estudo sobre a CMS Collaboration (CERN), que evidenciou a crescente prática da hipercoautoria, caracterizada por publicações com centenas ou até milhares de coautores. Entre 2014 e 2023, foram analisadas 1.932 publicações, das quais a maioria contou com mais de 400 autores, indicando a importância das colaborações em larga escala para enfrentar desafios científicos complexos. Essa dinâmica reforça a relevância da cooperação internacional na produção e disseminação do conhecimento, especialmente em áreas como a física de partículas.

De acordo com Zitt e Bassecoulard (2004), ao analisar a presença de autores internacionais em publicações científicas, bem como a diversidade geográfica das colaborações, é possível identificar padrões e tendências na globalização da produção científica. Segundo os autores, a internacionalização não se limita à coautoria, mas também se manifesta na circulação de ideias, na escolha de periódicos internacionais e na disseminação do conhecimento científico, o que sugere uma crescente interdependência entre países e instituições.

Sidone, Haddad e Mena-Chalco (2016) argumentam que a consideração do espaço geográfico na geração do saber, especialmente nas parcerias acadêmicas entre pesquisadores, permite compreender mais profundamente as redes de pesquisa e subsidiar estratégias e políticas voltadas à ciência, tecnologia e inovação no Brasil. Yan e Sugimoto (2011) investigaram os padrões de citações e as conexões colaborativas entre organizações do campo da Biblioteconomia e da CI, e concluíram que pesquisadores costumam referenciar colegas do mesmo país e/ou indivíduos geograficamente próximos. Isso é corroborado por Head, Li e Minondo (2018), que propõem a ideia de que a proximidade territorial contribui para o estabelecimento de laços pessoais, os quais, por sua vez, estimulam a disseminação do conhecimento.

Reis (2021) realizou uma pesquisa (Figura 5) com o objetivo de investigar a relação entre colaboração científica internacional e formação docente internacional (doutorado e pós-doutorado), por análise de coautoria da produção científica institucional.

Figura 5 – Principais etapas do procedimento experimental de Reis em 2021.



Fonte: Reis (2021).

A imagem apresenta o fluxo metodológico de integração entre dados da PL e da WoS para analisar o impacto da formação docente na colaboração e produção científica. Na Etapa 1, são recuperados 35.365 registros do Lattes, dos quais 1.915 currículos são selecionados com metadados de formação docente. Na Etapa 2, são obtidos 18.680 registros da WoS. Em seguida, ocorrem duas fusões: a primeira pela correspondência de DOI, e a segunda pelo conjunto ISSN, ano, volume e página inicial, resultando em 12.787 registros bibliográficos coincidentes entre Lattes e WoS. A Etapa 3 envolve o tratamento e padronização dos metadados. A Etapa 4 realiza a fusão final das bases por identificadores únicos. Por fim, a Etapa 5 possibilita criar indicadores sobre como a formação docente influencia a colaboração e a produção científica.

No entanto, Reis (2021) não focou na questão de endogamia acadêmica. Nesse sentido, seria possível expandir sua pesquisa comparando a colaboração internacional de indivíduos endógenos e não-endógenos, com e sem experiência no exterior (por ex., em um estágio de pós-doutorado), levando em consideração a definição de endogenia acadêmica de Berelson (1960), que consiste no recrutamento de um pesquisador pela mesma instituição que ele fez o doutorado. Por sua vez, Grochocki e Cabello (2022) analisaram a relação entre endogamia e mobilidade considerando o momento em que esta ocorre, distinguindo a mobilidade anterior à contratação e a adquirida ao longo da carreira, o que permitiu avaliar como diferentes graus e tipos de mobilidade (nacional e internacional) influenciam os resultados, tendo concluído que a endogamia acadêmica está associada à menor produtividade científica, sobretudo quando acompanhada de imobilidade profissional.

Nesse âmbito de fomento a experiência internacional, o material institucional do Ciência sem Fronteiras (CSF) relata que ele é “um programa que busca promover a consolidação, expansão e internacionalização da ciência e tecnologia, da inovação e da competitividade brasileira por meio do intercâmbio e da mobilidade internacional” (CSF, 2022).

Borenstein, Perlin e Imasato (2022) argumentam que a mobilidade dos acadêmicos no ensino superior brasileiro é muito baixa, pois não há incentivo para que os pesquisadores troquem de instituição ou se afastem do ambiente familiar. Os autores descobriram que pesquisadores endógenos que passaram por um período de mobilidade antes de ingressarem na instituição que fizeram o doutorado, são mais produtivos do que os pesquisadores endógenos que não tiveram experiência fora da sua instituição. Como continuação dessa pesquisa, poderia ser interessante comparar, proporcionalmente, a formação docente internacional de indivíduos não-endógenos e endógenos. A exploração das trajetórias de estudantes é vista como estratégica para mensurar ativos intangíveis, fortalecer o planejamento estratégico e ampliar o impacto da formação acadêmica, contribuindo para a geração de conhecimento e inovação social (Taxweiler; Sell; Pacheco, 2023).

De acordo com Braga e Venturini (2013), a elevada endogenia tem sido apontada como um fator negativo para a produtividade acadêmica e para a qualidade do conhecimento produzido. Horta (2013), por meio de análise da

literatura, inferiu que o endógeno acadêmico possui níveis mais baixos de desempenho relacionados a produtividade, com exceção de um estudo que foi feito pela Universidade do Texas. Praticamente uma década após esses trabalhos, os autores Borenstein, Perlin e Imasato (2022) investigaram o impacto da formação acadêmica endógena na produtividade de pesquisadores brasileiros, por meio de um volumoso conjunto de dados provenientes da PL e do Google Acadêmico. A pesquisa teve como principal objetivo avaliar se a endogenia tem um efeito prejudicial na produtividade da pesquisa no Brasil, levando em consideração diversas áreas do conhecimento. Como conclusão, foi identificado que pesquisadores endógenos são significativamente mais produtivos do que os não endógenos em todas as publicações de pesquisa, com exceção dos livros, ou seja, em relação a produtividade científica, para os autores não há provas que a endogenia acadêmica traz um efeito negativo.

Bastos, Zago e Recuero (2016) mensuraram a endogamia acadêmica na área da Comunicação, por meio da análise de redes sociais aplicada às coautorias e colaborações, utilizando dados da PL. Utilizando métricas como modularidade, grau de conexão e centralidade, os autores identificaram redes fragmentadas, com predomínio de vínculos institucionais e baixa interação entre grupos, revelando que as colaborações ocorrem quase exclusivamente dentro das mesmas instituições, caracterizando a endogamia acadêmica da área de Comunicação.

A respeito da pós-graduação no Brasil,

> Desde a sua implementação pelo governo militar, o maior desafio do sistema de ensino da pós-graduação no Brasil é a formação de recursos humanos a fim de capacitar os docentes das universidades, integrar a pós-graduação no sistema universitário, valorizar as ciências básicas e evitar disparidades regionais (Pelegrini; França, 2020, p. 574).

Maciel *et al*. (2018) argumentam que para uma adequada gestão dos PPGs, é necessário se apropriar de ferramentas para subsidiar o monitoramento das atividades que impactam a avaliação dos PPGs, principalmente no que diz respeito a produção científica. Segundo Damaceno, Haddad e Mena-Chalco (2018), a influência de uma instituição está ligada à formação prévia de seus pesquisadores, ou seja, aonde eles obtiveram os títulos de mestre e doutor. O estudo de Dias, Moita e Dias (2019) menciona a análise da endogamia acadêmica em universidades brasileiras, indicando que instituições com maior porcentagem de PPGs avaliados

com notas elevadas pela CAPES tendem a apresentar níveis menores de endogamia, sugerindo que a diversidade institucional está associada à maior qualidade acadêmica (Dias; Moita; Dias, 2019).

Alves, Faria e Amaral (2017) apontam que um corpo docente com formação mais diversificada contribui para a qualidade dos PPGs, uma vez que possibilita a incorporação de experiências obtidas em outras instituições. Essa informação é corroborada por um estudo divulgado pela revista digital da Fapesp[12], que foi publicado na revista Higher Education Quarterly, relatando que pesquisadores que se estabelecem em uma mesma instituição desde a sua formação (isto é, endógenos), tendem a realizar pesquisas de menor qualidade. Porém, em um cenário onde não existe diferença salarial entre instituições (por ex., entre as IES federais brasileiras), há pouca motivação para a mobilidade acadêmica por partes dos pesquisadores, tornando o escopo da seleção mais local (Horta; Yudkevich, 2016), sendo comum o fato dos pesquisadores completarem todas as etapas de formação dentro da mesma região, mostrando uma baixa mobilidade dos cientistas brasileiros (Furtado *et al*., 2015).

Horta (2013), sugere uma taxonomia de modo a distinguir os tipos de endogenia acadêmica, podendo ser vista no Quadro 1.

Quadro 1 – Taxonomia das categorias de endógenos.

Categoria	Descrição
Endógeno puro	Imóvel, passou toda a carreira na mesma universidade
Endógeno móvel	Esteve em outra universidade durante o doutorado (“sanduíche”) ou fez pós-doutorado em outra universidade
Cordão de prata	Trabalha no local onde realizou o doutorado, porém, após a conclusão do doutorado iniciou a carreira acadêmica em outro lugar

Fonte: Adaptado de Horta (2013).

A endogenia acadêmica pode ser observada em diversos contextos, como:

1. Endogenia em periódicos científicos:

[12] Revista Fapesp. Disponível em: <https://revistapesquisa.fapesp.br/horizonte-limitado>. Acesso em: 18 jul. 2024.

a. Endogenia de instituição, que ocorre quando há priorização na escolha de trabalhos de pesquisadores que pertencem a mesma instituição do editor do periódico (Amoras, 2017);

b. Endogenia de unidade da federação (estado): no contexto de área da CAPES, é calculada a partir da quantidade de autores, pareceristas e integrantes do conselho editorial que atuam na mesma unidade da federação da instituição responsável pelo periódico (Amoras, 2017).

2. Endogenia em uma instituição, que pode ser motivada por vínculos anteriores (*networking*), como: afiliados do laboratório onde realizaram o doutorado (Shibayama, 2022); publicações em coautoria (Rocca, 2007); etc.

3. Endogenia em uma unidade da federação: como exemplo, há uma pesquisa realizada em 2021 pelo Centro de Gestão e Estudos Estratégicos (CGEE), que é uma organização social brasileira vinculada ao MCTI, que analisou mestres e doutores, tendo como referência o grau de endogenia nos anos de 2009 e 2017, considerando duas variáveis, onde

> no caso que toma como referência o ano de 2009, a primeira variável é o número de mestres ou doutores titulados em uma determinada unidade da federação (no período 1996-2009), que estavam empregados na mesma unidade da federação no dia 31/12/2009. [...] a segunda variável é o número total de mestres ou doutores titulados em todas as unidades da federação no mesmo período (1996-2009) que se encontravam empregados na referida unidade da federação naquela mesma data. O resultado da divisão da primeira variável pela segunda variável (medida em termos percentuais) é chamado de grau de endogenia. Isso foi feito para o ano de 2017, com o período de titulação de 1996 a 2017 e o emprego em 31/12/2017 (CGEE, 2021).

Em relação às questões geográficas, as quais utilizam os conceitos de geolocalização (ou georreferenciação), que é o processo que possibilita localizar geograficamente determinado objeto espacial, num sistema de referência, por meio de coordenadas, Furtado *et al*. (2015) observaram que: apenas 20% dos pesquisadores brasileiros trabalham a mais de 500 km da instituição onde começaram a trajetória acadêmica; a maioria dos pesquisadores trabalham a menos de 100 km da universidade onde iniciaram a carreira; entre os pesquisadores que

fizeram doutorado ou pós-doutorado no exterior, 81% se estabeleceram em suas regiões de origem.

Segundo Boschma (2005), a capacidade das organizações de adquirir conhecimento e desenvolver inovações pode depender da proximidade social, no entanto, ela em excesso pode gerar efeitos negativos sobre o aprendizado e a inovação. Ainda de acordo com o autor, proximidade institucional em demasia é desfavorável para o surgimento de novas ideias e inovações devido ao enrijecimento institucional (que dificulta a percepção de novas oportunidades) e à inércia (que atrapalha os reajustes institucionais necessários). Por outro lado, uma proximidade institucional muito limitada compromete a ação coletiva e o processo inovador em razão da fragilidade das instituições formais e da ausência de coesão social e valores compartilhados (Boschma, 2005).

Fazendo um paralelo com o conceito de redes, mesmo uma ligação fraca tem o seu valor, pois “sistemas sociais que não contenham ligações fracas serão fragmentados e incoerentes, o conhecimento científico será prejudicado e haverá a formação de subgrupos raciais, étnicos ou geográficos” (Braga; Gomes; Ruediger, 2008).

Para Damaceno, Haddad e Mena-Chalco (2018), há uma ausência de estudos sobre a influência da distribuição geográfica dos pesquisadores das IES, provavelmente, entre outros fatores, devido à baixa disponibilidade e qualidade de bases de dados com informações a respeito da instituição de trabalho do profissional.

Borenstein, Perlin e Imasato (2022) consideram que apenas uma pesquisa crítica dos efeitos de endogenia acadêmica pode auxiliar os políticos a analisaram os seus impactos negativos e positivos no sistema acadêmico brasileiro, ao qual passa por um momento de grande pressão por parte da sociedade, ou seja, ela necessita ser melhor explorada e discutida.

No contexto científico, onde o volume de publicações, citações e dados de colaboração cresce exponencialmente, a área de visualização de informação é uma ferramenta analítica indispensável. Brinton (1939) relatou que William Playfair desenvolveu os seus primeiros gráficos a partir do ano de 1786, ou seja, a informação representada de maneira visual tem sido estudada há séculos.

A visualização de informação estuda a utilização de interfaces interativas, com seus respectivos filtros e gráficos, para a representação visual de dados.

> O interesse sobre a maneira como artefatos visuais e interativos podem ser desenvolvidos visando auxiliar esse processo de cognição [...] deu origem à área de pesquisa denominada Visualização de Informação (*Information Visualization*), que estuda o uso de representações visuais e interativas de dados abstratos e não baseados em aspectos físicos, com o propósito de ampliar a cognição (Card; Mackinlay; Shneiderman, 1999).

Para Ribeiro (2012), a análise de gráficos não é trivial, além disso, alguns elementos interferem na compreensão, como: a intimidade da pessoa com a temática apresentada; a aparência da visualização; e a quantidade de informações em um gráfico.

A partir do trabalho de Freitas *et al*. (2001), é possível inferir que a visualização de informação pode ser entendida como o estudo das principais formas de representações gráficas para apresentação de informações, utilizando-se de aspectos de computação gráfica, interfaces homem-computador e mineração de dados, com o objetivo de potencializar a compreensão da informação, além de auxiliar o usuário na criação de novos conhecimentos.

De acordo com Alves (2015), a área de visualização de informação pode auxiliar no entendimento dos resultados alcançados pelos pesquisadores brasileiros. Reis (2016) corrobora com esse pensamento, pois, segundo o autor, por mais que existem vários estudos sobre a elaboração de indicadores bibliométricos por meio da análise da produção científica que é indexada em bases de dados, é necessário desenvolver soluções tecnológicas que possibilitem a visualização desses indicadores, de modo a ampliar a compreensão a respeito dos resultados das atividades de pesquisa. Porém, Alves, Faria e Amaral (2017) argumentam que a visualização de informação é pouco explorada no contexto de indicadores acadêmicos.

Reis (2016) menciona que a área de desenvolvimento de *software* tem contribuído para a visualização de informação, já que vem sendo desenvolvidos vários componentes informatizados, como *plug-ins*, extensões e APIs, que possuem o objetivo de adicionar recursos aos sistemas informatizados que utilizam conjuntos de dados volumosos. As APIs possuem documentações técnicas que contêm

informações detalhadas a respeito da estrutura e do acesso aos dados (Rodrigues, 2024).

Uma vez contextualizada a importância da visualização de informação, deve ser ponderado como os dados estão sendo representados. Silva, Gomes e Rodrigues (2023) argumentam que atributos relacionados a *User Interface* (UI) interferem na disponibilidade e recuperação de informação.

A respeito de indexação, pode ser entendida como uma forma de escolha dos termos que representam de forma mais apropriada um determinado documento (Lima, 2020). Baptista e Machado (2001) apontam que por mais que haja uma alta tecnologia nos sistemas de indexação e de recuperação da informação, a relevância e a precisão das respostas que eles entregam aos usuários ainda não atingiram níveis satisfatórios, logo, é necessário melhorar a eficácia e a eficiência dos serviços de informação. Para tanto, foram criados os chamados metadados semânticos.

De acordo com Angelozzi e Martín (2010), distintos esquemas de metadados irão coexistir e cada um deles poderá ser utilizado para um fim, portanto, é necessário refletir sobre a questão da interoperabilidade, a qual permite a troca de dados entre esquemas.

Segundo Baptista (2007), o debate a respeito de metadados deve considerar:

- Aspectos conceituais e tecnológicos;
- A diversidade de objetos que são convertidos em recursos informacionais;
- Os diferentes atores que participam do fluxo informacional, tais como: produtores de informação, desenvolvedores de *software*, profissionais da informação, agências normativas e os usuários da informação.

Para Brito e Martínez-Ávila (2019), os metadados tem a função de padronização e integração dos sistemas de informação, para que seja possível trocar e compartilhar informações, estando relacionados à catalogação de recursos informacionais no meio digital, descrevendo e representando os documentos.

Uma notação de metadados que vem sendo largamente utilizada no desenvolvimento de sistemas para Web é chamada de JSON[13] e se constitui em um formato leve de troca de dados, ou seja, ele permite a interoperabilidade entre sistemas.

A Figura 6 mostra um comparativo entre os padrões JSON e XML.

Figura 6 – Comparativo entre as estruturas dos padrões XML e JSON.

Estrutura em XML	<?xml version="1.0" encoding="UTF-8"?> <livro> <titulo>JavaScript: Guia do programador</titulo> <autor>Maurício Samy Silva</autor> <ano>2010</ano> </livro>

Estrutura equivalente em JSON	{
	"titulo": "JavaScript: Guia do programador",
	"autor": "Maurício Samy Silva",
	"ano": 2010
	}

Fonte: elaborada pelo autor.

Conforme pode ser visto, foi estabelecido um comparativo entre os padrões XML e JSON, mostrando como a mesma informação pode ser representada por sintaxes distintas. Na parte superior, o XML utiliza uma estrutura baseada em *tags* de marcação (como `<autor>` e `</autor>`), o que confere uma hierarquia explícita e visualmente segmentada, porém mais volumosa. Em contrapartida, a parte inferior exibe a estrutura equivalente em JSON, que simplifica a representação ao adotar pares de chave-valor delimitados por chaves {}. Enquanto o XML exige a repetição dos rótulos para fechar cada campo, o JSON elimina essa redundância, apresentando os dados de forma mais compacta e direta, mantendo a integridade das informações originais: título, autor e ano de publicação.

Para Arakaki, Simionato e Santos (2017), as tecnologias e os padrões que são usados no campo da informação possuem um importante papel nos processos sociais e econômicos, criando oportunidades aos profissionais que possuem

[13] JSON. Disponível em: <https://www.json.org>. Acesso em: 18 jul. 2024.

competências a respeito de tratamento e representação da informação. Ribeiro (2012) argumenta que a visualização de informação é tema de pesquisas e de atuação profissional em várias instituições no mundo, sendo o jornalismo uma das áreas profissionais que se apropriou dos conhecimentos produzidos em visualização de informação, fornecendo recursos visuais que impactam o dia a dia das pessoas.

Patil (2011) sugere que o aumento na procura por cientistas de dados foi impulsionado pelo sucesso das grandes empresas de Internet, locais onde esses profissionais podem desempenhar vários papéis. Segundo Sarvo, Reis e Amaral (2022), uma possível solução para o mapeamento das atividades das IES é o gerenciamento eficiente de dados institucionais, amparado por sistemas de informação interoperáveis, que usam fontes de informações internas e externas com dados a respeito da instituição.

Nesse contexto de Ciência de Dados, há um conceito muito utilizado, chamado de *big data*, que

> se refere a dados que são grandes demais para um único servidor, muito diversos para se adequar a uma base de dados estruturada em linhas e colunas, ou cujo fluxo seja tão intenso que não permita adequação a um data warehouse estático (Davenport, 2014).

Para Camperos-Reyes *et al*. (2019), “o fenômeno *big data* se refere a geração e uso de dados estruturados, semi-estruturados, em que têm sido adotadas diversas técnicas para coletar, integrar, analisar e tomar decisões, a partir de diversos tipos de dados, para os mais diversos fins”.

Guimarães, Rocha e Mugnaini (2023) apresentam um estudo cientométrico que analisa a produção acadêmica relacionada às temáticas de humanidades digitais e *big data* nas universidades estaduais paulistas — UNESP, UNICAMP e USP — entre 2006 e 2021, tendo por objetivo principal mapear e caracterizar a evolução dessas áreas, considerando sua inserção no contexto da ciência aberta e da interdisciplinaridade. Utilizando dados da base WoS e análise por coocorrência de palavras-chave, os autores identificaram um crescimento progressivo no número de publicações, especialmente a partir de 2015, com destaque para a USP. As produções evidenciam a ampliação do interesse acadêmico pelos impactos das tecnologias digitais nas ciências humanas e sociais, bem como a crescente

aplicação de métodos computacionais nessas áreas (Guimarães; Rocha; Mugnaini, 2023).

De acordo com Dumbill (2012), no âmbito de caracterizar os diferentes aspectos de *big data*, há os chamados 3 Vs, que representam Volume, Velocidade e Variedade, onde a partir deles, torna-se mais fácil compreender a natureza dos dados. A seguir, cada um dos Vs é descrito (Dumbill, 2012):

- Volume: diz respeito ao benefício obtido com a capacidade de processar grandes quantidades de informações, embora sejam requeridos armazenamento escalável e abordagem distribuída para consultas;
- Velocidade: esse aspecto está relacionado ao fluxo crescente com que os dados se disseminam em uma organização;
- Variedade: uma vez que dificilmente os dados de origem se apresentarão de modo organizado, é convencional que em sistemas de *big data* o conjunto de dados não seja estruturado de forma relacional.

Segundo Ottonicar, Atayde e Santa-Eulalia (2019), o ato de analisar dados (e metadados) que estão disponíveis em sistemas informatizados contribui para entender as tendências do mercado, possibilitando aos gestores desenvolverem oportunidades de negócios. Para os autores, os grandes conjuntos de dados precisam ser armazenados de forma segura, respeitando aspectos como privacidade e ética no uso de dados pessoais e organizacionais. Isso vai ao encontro do que é preconizado pela Lei Brasileira de Proteção de Dados Pessoais (LGPD)[14] e por autores como Zimmer e Proferes (2014), que destacam a importância de anonimizar dados no contexto de *big data*. Nesse domínio, Hoeren e Kolany-Raiser (2017) apresentam o princípio da limitação da finalidade, o qual estabelece que os dados pessoais só podem ser obtidos para uma determinada finalidade previamente definida, que seja clara e lícita.

Para Zhu e Xiong (2015), *big data* é um dos principais tópicos de pesquisa em ciência de dados no contexto da indústria, podendo beneficiar investigações

[14] LGPD. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>. Acesso em: 18 jul. 2024.

científicas (Dhar, 2013). A ciência de dados trabalha com métodos e técnicas que incluem obtenção, armazenamento, gerenciamento, segurança, análise e visualização de dados. Ressalta-se que a computação em nuvem auxilia a adoção de *big data*, uma vez que não é necessário um grande investimento inicial na aquisição de *hardware* para realizar estudos com grandes conjuntos de dados.

Sant'Ana (2016) aponta que o custo de aluguel e manutenção de serviços digitais está cada vez mais acessível. Isso quer dizer que o *hardware* compartilhado nas arquiteturas de nuvem coloca o processamento de grandes conjuntos de dados ao alcance de pequenas empresas e de pessoas físicas, que podem usar tempo de servidor na nuvem, otimizando o investimento orçamentário.

Superada a barreira dos custos de infraestrutura computacional, deve-se analisar os tipos de dados que serão processados. Yamaguchi (2010) indica que antes de utilizar uma visualização deve-se identificar o tipo de dado que será representado. A autora desenvolveu um sumário de caracterização de dados, tendo como referência os trabalhos de Keim (2002), Freitas *et al*. (2001) e Shneiderman (1996), conforme pode ser visto no Quadro 2.

Quadro 2 – Caracterização dos tipos de dados.

Critério	Classe	Exemplos
Classe de informação	Texto/Web	Documentos: pdf, doc, html
	Hierárquico e grafos	Organograma, redes
	Algoritmos e *softwares*	Código-fonte, logs
Natureza do domínio	Qualitativos nominais	Gênero, estado civil, nacionalidade
	Qualitativos ordinais	Nível: básico, intermediário, avançado
	Quantitativos discretos	Quantidade de vendas
	Quantitativos contínuos	Renda per capita
Dimensão	1-D	Dados temporais: data hora, hora, intervalo de tempo
	2-D	Dados geográficos: mapas, superfícies de terrenos
	3-D	Dados de objetos do cotidiano
	n-D	Tabelas de base de dados

Fonte: Adaptado de Yamaguchi (2010).

Como pode ser visto, é apresentada uma caracterização que ajuda a identificar o tipo de dado e sua estrutura, orientando escolhas metodológicas para análises.

Na classe de informação, os dados podem assumir formatos variados, como textos e páginas da Web, estruturas hierárquicas ou grafos, além de algoritmos e *softwares*. A natureza do domínio descreve o tipo de variável representada. Os dados podem ser qualitativos nominais, que não têm ordem natural, qualitativos ordinais, que possuem hierarquia ou nível, quantitativos discretos, que só podem assumir valores inteiros, normalmente resultantes de contagem, ou quantitativos contínuos, que assumem valores em intervalos, ou seja, não se limitam a números inteiros — podem assumir qualquer valor dentro de uma faixa, incluindo números fracionários. Por fim, o critério da dimensão indica a complexidade do dado.

Para Few (2006), os elementos visuais disponíveis na ferramenta gráfica de *software* conhecido como *dashboard* ajudam a visão e o cérebro humano no processo de obtenção de informações importantes. Sedrakyan, Mannens e Verbert (2019) mencionam que há vários provedores de *dashboards*, porém, há de se tomar cuidado com suas visualizações, pois podem gerar interpretações errôneas.

O Quadro 3 apresenta algumas características de *dashboard*, com suas variáveis e valores.

Quadro 3 – Categorização de *dashboard*.

Variável	**Valor**
Papel	Estratégico Analítico Operacional
Domínio dos dados	Marketing Produção Recursos Humanos
Alcance dos dados	Toda a empresa Departamental Individual
Frequência de atualização	Mensalmente Semanalmente Diariamente Em tempo real
Interatividade	Interface estática Interface dinâmica

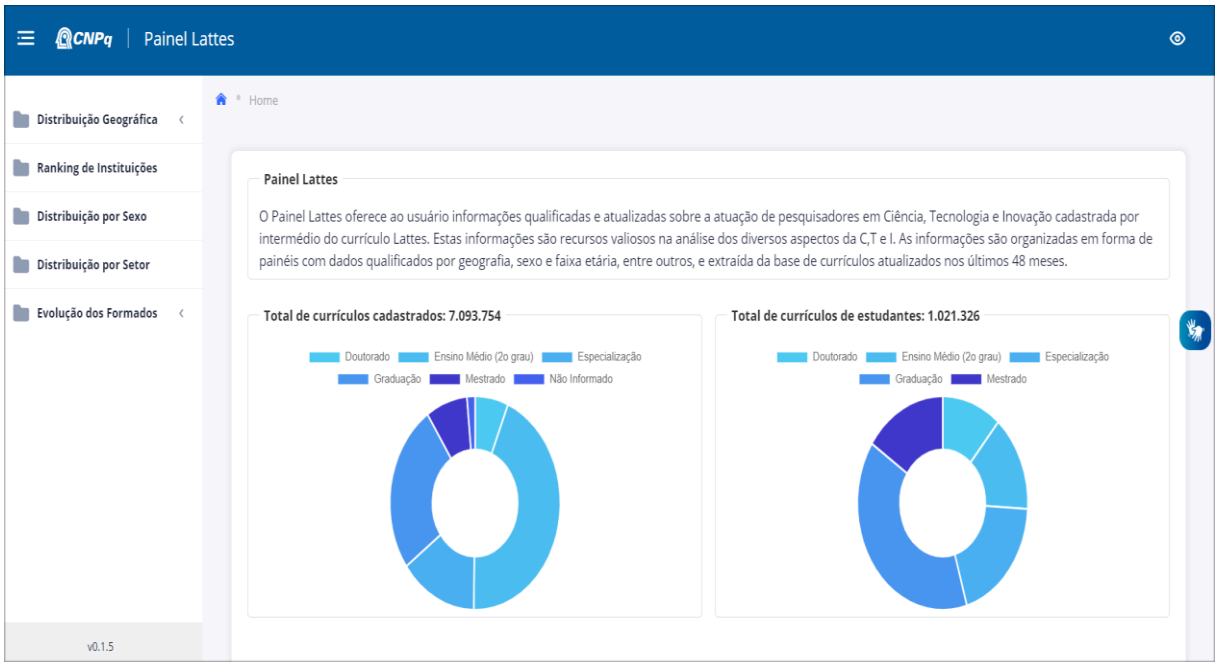
Fonte: Adaptado de Few, 2006.

A primeira variável, Papel, indica o nível e a função do sistema. A variável Domínio dos dados especifica em qual área organizacional os dados do sistema estão concentrados. Em seguida, a variável Alcance dos dados descreve o nível de abrangência das informações manipuladas. A variável Frequência de atualização mostra a periodicidade com que os dados são renovados. Por fim, a variável Interatividade indica o tipo de interface oferecida ao usuário, podendo ser estática — limitada a consultas simples — ou dinâmica, permitindo maior interação, filtragem e manipulação dos dados.

Em relação aos *dashboards* de dados acadêmicos e científicos, como os disponibilizados por ICTs, alguns casos merecem destaque e serão apresentados a seguir.

Como pode ser visto na Figura 7, o CNPq, por meio do Painel Lattes[15], disponibiliza um *dashboard* com indicadores provenientes dos CVs Lattes.

Figura 7 – Painel Lattes.



Fonte: Painel Lattes, acesso realizado pelo autor.

Como pode ser notado, o Painel Lattes fornece informações sobre a atuação de pesquisadores em CT&I a partir dos dados inseridos por eles nos CVs Lattes.

[15] Painel Lattes. Disponível em: <https://painel-lattes.cnpq.br>. Acesso em: 18 jul. 2024.

Outro exemplo de *dashboard* é o BrCris (Figura 8), criado pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), que “é uma plataforma computacional para integração, visualização e prospecção de dados científicos com a finalidade de estabelecer um modelo único de organização da informação científica de todo o ecossistema da pesquisa brasileira” (IBICT, 2023).

Figura 8 – BrCris.

Fonte: BrCris, acesso realizado pelo autor.

Como pode ser observado, o BrCris possui uma ferramenta de busca onde é possível consultar informações a respeito de publicações, pessoas, revistas, organizações, patentes, PPGs, grupos de pesquisa e *softwares*. “É uma plataforma agregadora que permite recuperar, certificar e visualizar dados e informações relativas aos diversos atores que atuam na pesquisa científica do contexto brasileiro” (IBICT, 2023). Além disso, o BrCris disponibiliza visualização de redes de colaboração e painéis de indicadores em CT&I.

É importante ressaltar que, conforme pode ser visto na Figura 9, o Observatório de CT&I em Saúde da Fiocruz serviu como projeto piloto do BrCris (Fiocruz, 2023).

Figura 9 – Observatório da Fiocruz.

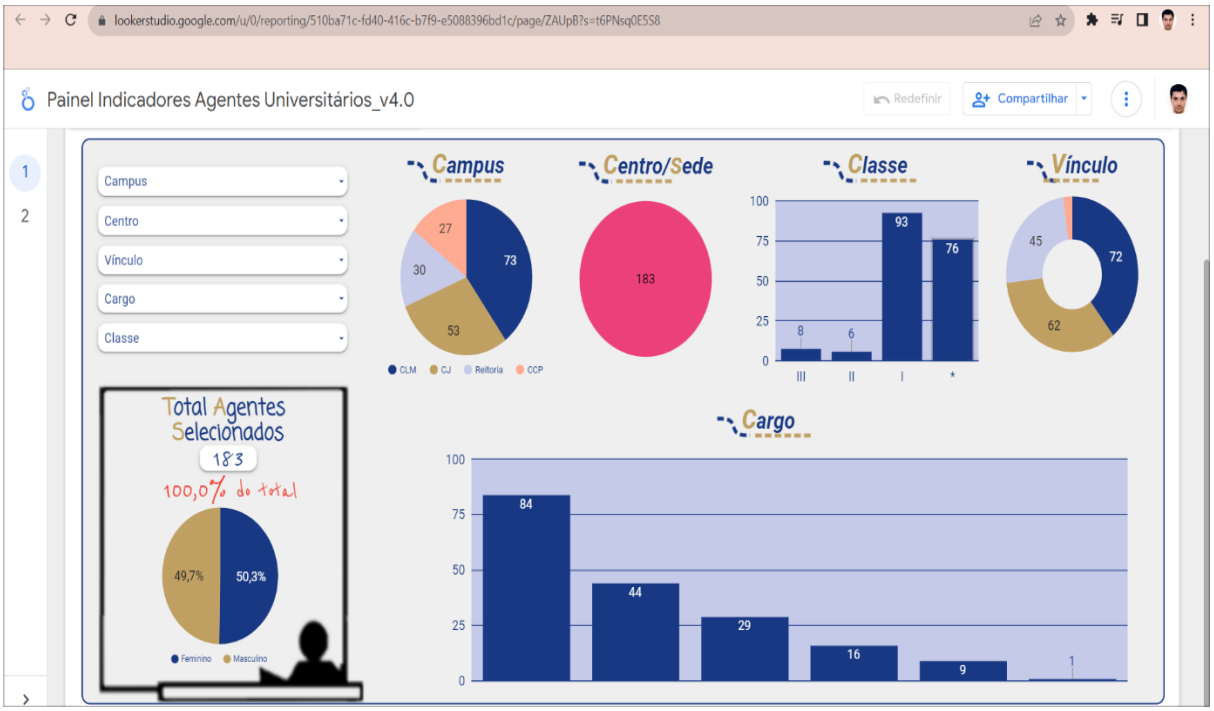
Fonte: Observatório da Fiocruz, acesso realizado pelo autor.

“Além da disponibilidade de bases de dados e fontes de dados específicas, o trabalho cientométrico normalmente também requer ferramentas e instrumentos adicionais, necessários para a análise adequada de dados e indicadores derivados” (Costas, 2017). Nesse âmbito de aparatos tecnológicos, Norman (2006) argumenta que um produto com o *design* mal construído durante a fase de projeto, ocasionará dificuldade de utilização por parte dos usuários. Por sua vez, o Google possui uma ferramenta chamada Data Studio[16], que possibilita a criação de visualizações gráficas por meio de uma interface amigável, permitindo integrar com outros recursos do ecossistema dele, como Analytics e Google Ads.

Em relação às plataformas científicas que utilizam o Google Data Studio, existe uma iniciativa da Universidade Estadual do Norte do Paraná (UENP) chamada UENP em números[17], onde, na Figura 10, é mostrado um dos painéis dessa aplicação.

[16] Google Data Studio. Disponível em: <https://datastudio.withgoogle.com>. Acesso em: 18 jul. 2024.
[17] UENP em números. Disponível em: <https://uenp.edu.br/uenp-dados>. Acesso em: 18 jul. 2024.

Figura 10 – UENP em números.



Fonte: UENP, acesso realizado pelo autor.

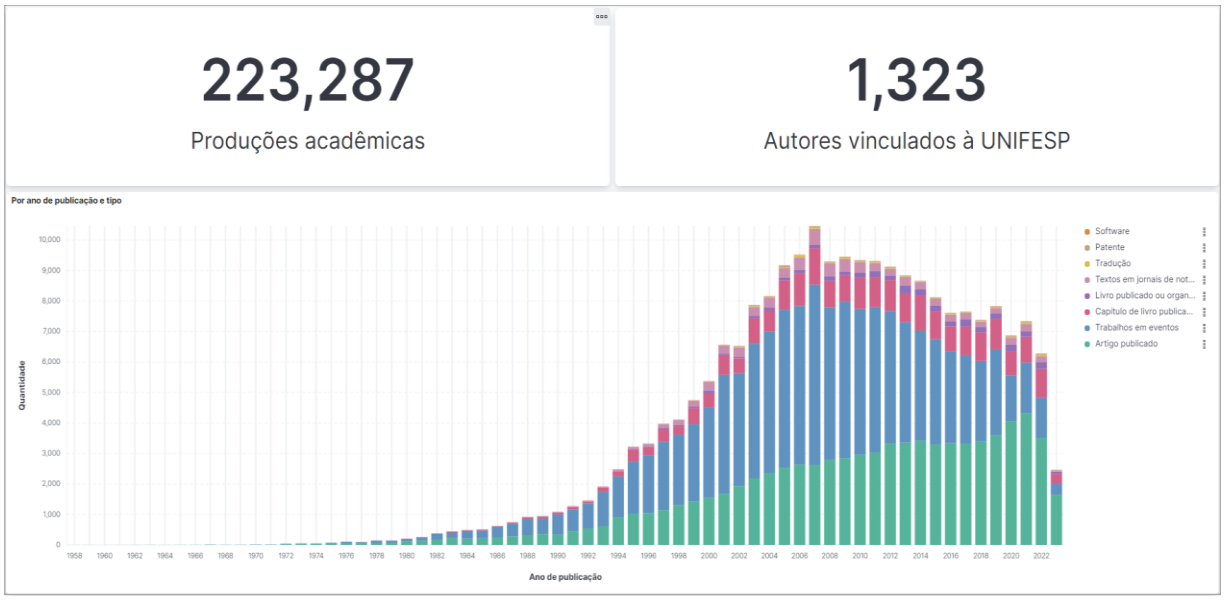
Conforme pode ser visto, o painel “Indicadores de Agentes Universitários” fornece aos usuários vários tipos de visualizações gráficas com os respectivos filtros, ou seja, é uma interface interativa.

Sobre tecnologias utilizadas no armazenamento dos conjuntos de dados utilizados em *dashboards*, é interessante comentar que a plataforma Prodmais, que foi apresentada anteriormente, utiliza a ferramenta Elastic Search, que, assim como o MongoDB, é um banco de dados orientado a documentos desnormalizados, possibilitando melhor desempenho na recuperação de dados, uma vez que nenhuma junção de consulta é necessária.

Já no painel Prodmais, é possível consultar duas categorias de *dashboard*: das produções acadêmicas e dos perfis dos pesquisadores, conforme pode ser visto a seguir.

A Figura 11 apresenta o *dashboard* das produções acadêmicas.

Figura 11 – Prodmais - *dashboard* das produções acadêmicas.

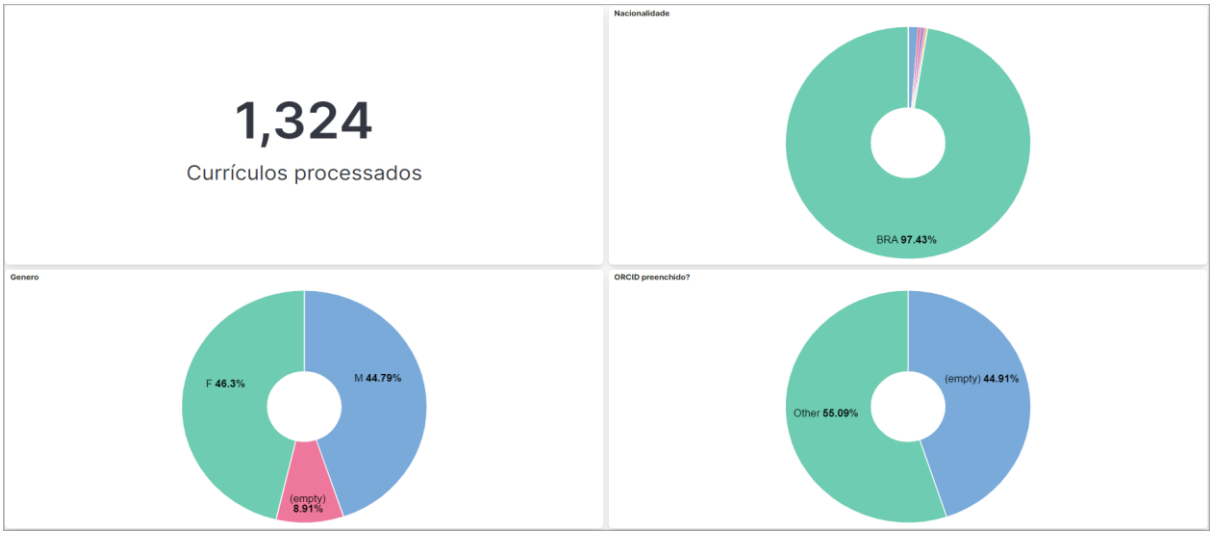


Fonte: Prodmais, busca realizada pelo autor.

Como pode ser observado, no presente momento há mais de 223 mil produções acadêmicas relacionadas aos 1323 autores que possuem vínculo com a UNIFESP.

A Figura 12 mostra o *dashboard* dos perfis dos pesquisadores.

Figura 12 – Prodmais - *dashboard* dos perfis dos pesquisadores.



Fonte: Prodmais, busca realizada pelo autor.

O *dashboard* dos perfis dos pesquisadores mostra, dentre outras informações, que mais de 97% são de nacionalidade brasileira. Um detalhe

importante que pode ser visto no gráfico de pizza sobre gênero diz respeito ao dado "*empty*" (vazio), correspondendo a 8,91%. Dependendo da análise que é feita, esse tipo de dado necessita ser retirado do recorte para não gerar interpretações errôneas. Como exemplo hipotético, se considerarmos que deveriam haver apenas dois possíveis valores referentes a esse dado na amostra (sexo feminino e sexo masculino), se olharmos somente para a informação que 46,3% são mulheres, poderíamos então inferir que a maioria é homem, o que no caso está errado, pois a quantidade de homens nesse exemplo hipotético (que considera apenas dois possíveis valores), representa 44,79% da amostra. Porém, não há problemas se for proposital a existência desse dado vazio. Esse foi apenas um exemplo a respeito do cuidado que devemos ter na interpretação de indicadores.

Dumbill (2012) argumenta que compreender a natureza do problema de *big data* é o primeiro passo na avaliação de soluções. Para Zhu e Xiong (2015), a utilização de um grande volume de dados para resolver problemas nas áreas científicas e sociais também fazem parte da ciência de dados.

De acordo com Camperos-Reyes *et al*. (2019), a geração e o processamento de dados têm influência sobre setores de grande importância para o desenvolvimento social e econômico. Ainda segundo os autores, os PPGs, que são mantidos por instituições de ensino e pesquisa, são relevantes meios para a formação de mão de obra qualificada.

Damaceno, Haddad e Mena-Chalco (2018), argumentam que a disponibilidade de profissionais com educação formal nos níveis mais altos (mestrado e doutorado) é um indicador das condições de prosperidade de um país, conhecido como o capital humano de uma economia. "A expansão e consolidação da pós-graduação stricto sensu são fundamentais para garantir a formação de pessoal qualificado para atuar no setor produtivo e nas universidades e promover o desenvolvimento socioeconômico do país" (Oliveira; Amaral, 2017).

Segundo Wyatt *et al*. (2017), por ser um campo de estudos interdisciplinar, CTS se apropria de uma ampla variedade de métodos, como observação, entrevistas e leitura de materiais, porém, métodos quantitativos baseados em dados numéricos de pesquisas em grande escala e visualizações de dados são menos comuns no CTS. Vale ressaltar que o Programa de Pós-Graduação em Ciência,

Tecnologia e Sociedade da UFSCar (PPGCTS/UFSCar)[18], além de interdisciplinar, conduz pesquisas teóricas e aplicadas em suas linhas de pesquisas, muitas vezes promovendo a inovação.

Para Oliva *et al*. (2023), o enfrentamento dos desafios urbanos — como crescimento populacional, mudanças climáticas, mobilidade e infraestrutura — depende da integração entre diversas áreas do conhecimento, como física, química, biologia, ciências sociais e matemática. Segundo os autores, a pesquisa científica, aliada à inovação tecnológica, permite a formulação de soluções práticas e embasadas, que influenciam diretamente o planejamento urbano e a melhoria da qualidade de vida.

Nesse contexto, o Ripoli *et al*. (2024) relatam a criação da InfoHub, uma plataforma tecnológica idealizada por uma equipe multidisciplinar no âmbito de um MBA da UFSCar, voltada à conexão entre projetos de pesquisa e instituições financiadoras. A proposta busca suprir a ausência, no Brasil, de uma solução digital unificada que promova inovação social por meio da integração entre ciência, tecnologia e políticas públicas. Utilizando ferramentas de inteligência artificial e ciência de dados, a InfoHub visa facilitar o encontro entre demandas do setor público e privado e iniciativas acadêmicas, promovendo colaboração interdisciplinar, eficiência e sustentabilidade nos processos de inovação.

Oliveira e Amaral (2017) argumentam que as políticas nacionais de educação e de ciência e tecnologia possibilitam a evolução da pós-graduação e da produção científica brasileira. Por sua vez, Alves, Faria e Amaral (2017) apontam que organizar dados na forma de indicadores, juntamente com o uso de conceitos da área de visualização de informação, pode auxiliar em decisões relacionadas na implementação de políticas de ciência e tecnologia.

[18] PPGCTS/UFSCar. Disponível em: <http://www.ppgcts.ufscar.br>. Acesso em: 07 fev. 2026.

3 PROCEDIMENTO METODOLÓGICO

A pesquisa foi realizada no NIT-Materiais/UFSCar, que é uma unidade do Departamento de Engenharia de Materiais vinculada ao departamento de CI, reconhecida como um programa de extensão da UFSCar.

Trata-se de uma abordagem quantitativa, sendo as fontes de dados: a) lista contendo os IDs Lattes dos pesquisadores credenciados em PPGs da UFSCar (todos os campi), proveniente da Plataforma Sucupira; b) PL (metadados de produção científica e CVs Lattes). As técnicas de análise utilizadas, envolvem, principalmente: *Data Science*; Bibliometria; Ferramentas automatizadas.

A pesquisa iniciou com o objetivo de compreender melhor o conceito de endogamia acadêmica. Após, foram implementados *scripts* para extrair dados do LinkedIn. Como o uso do LinkedIn mostrou-se inviável (conforme será explorado a seguir), foram realizadas extrações de dados da PL por meio do uso das ferramentas csv_lattes e synclattes. Em seguida, foram estudados sistemas gerenciadores de bancos de dados não-relacionais, como o MongoDB, devido, entre outras características, sua robustez, facilidade de uso, ampla documentação e por possuir versão gratuita.

Os dados extraídos da PL foram tratados para que os indicadores de endogamia fossem analisados. Nesse sentido, esta pesquisa é experimental e exploratória (Gil, 2017).

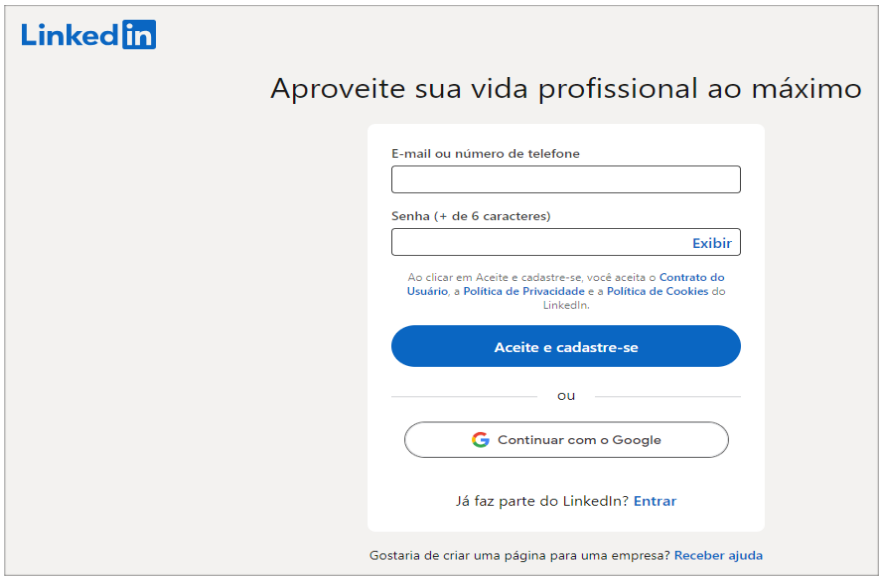
A ideia inicial desta pesquisa era utilizar o LinkedIn[19] como fonte de dados, que é uma rede social online, fundada em 2003 por Reid Hoffman e com sede em Mountain View (Califórnia).

Segundo informações do próprio LinkedIn, ele se constitui como a maior rede profissional do mundo, com mais de 850 milhões de usuários de 200 países e regiões.

A Figura 13 apresenta a tela de login do LinkedIn (site).

[19] LinkedIn. Disponível em: <https://about.linkedin.com/pt-br>. Acesso em: 18 jul. 2024.

Figura 13 – Tela de login do LinkedIn.



Fonte: LinkedIn, acesso realizado pelo autor.

Técnicas de extração de dados da Web, também conhecidas como *web scrapping* (raspagem de dados) e *crawler* (robôs rastreadores), possibilitam reunir uma grande quantidade de dados estruturados que são gerados pelos usuários.

Na literatura há trabalhos que exploraram as redes sociais. Podemos citar dois exemplos recentes:

- A pesquisa de Rodrigues e Sant'Ana (2023) sobre as APIs do Facebook, Twitter e LinkedIn;
- O estudo de Alzamora *et al*. (2022), que utilizou o Twitter para correlacionar o vocabulário das postagens com o agravamento e a atenuação da pandemia no Brasil.

Na presente pesquisa, optou-se por empregar a técnica de *web scrapping* do LinkedIn por meio de um *crawler*, uma vez que essa estratégia possibilita uma maior autonomia a respeito da escolha dos dados a serem extraídos.

Em dezembro de 2022 havia cerca de 50 mil ex-alunos da UFSCar registrados no LinkedIn na página UFSCar - Alumni[20], que é uma página não oficial,

[20] Página UFSCar - Alumni no LinkedIn. Disponível em: <https://www.linkedin.com/school/ufscar-alumini>. Acesso em: 18 jul. 2024.

criada e gerida por ex-alunos. Nessa página estão as pessoas que tiveram pelo menos em algum momento de suas vidas realizado algum curso na UFSCar e cadastrado essa informação no formulário de usuário do LinkedIn.

Importante lembrar que, até onde se sabe, não há auditoria pelo LinkedIn dos dados constantes nos perfis dos usuários, de modo que valide aquilo que foi preenchido, ou seja, não se tem controle a respeito do conteúdo que é inserido, apesar do LinkedIn citar em seus termos de uso que o usuário deve concordar em não publicar informações incorretas no perfil.

A Figura 14 mostra a página UFSCar - Alumni no LinkedIn.

Figura 14 – Página UFSCar - Alumni no LinkedIn.

Fonte: LinkedIn, acesso realizado pelo autor.

Em cada perfil de LinkedIn de um ex-aluno contém, dentre outros dados, o histórico profissional e o histórico acadêmico dele.

Há algumas formas de recuperar os perfis dos ex-alunos da UFSCar no LinkedIn. A primeira tentativa foi por meio do uso do motor de busca do Google[21]. A Tabela 1 mostra a expressão que foi utilizada.

[21] Google. Disponível em: <https://www.google.com.br>. Acesso em: 18 jul. 2024.

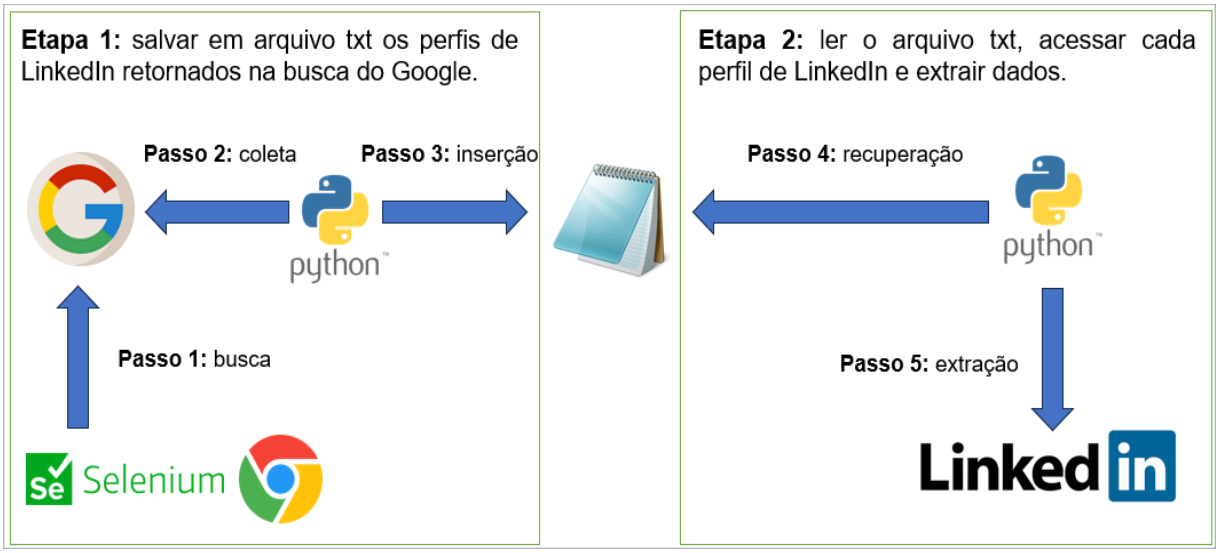
Tabela 1: Busca no Google.

Expressão	Data	Registros	Páginas
`site:linkedin.com/in/ AND "UFSCar"`	08/10/2022	~28.000	30 (10 perfis/pág)

Fonte: elaborada pelo autor.

Por sua vez, na Figura 15 é apresentado o processo que foi planejado, onde, por meio da linguagem de programação Python, do navegador Chrome e da ferramenta Selenium[22] (que serve para automatizar ações em aplicações Web), foi possível recuperar no Google os perfis de LinkedIn de alguns ex-alunos da UFSCar, visando armazenar a URL de cada perfil, para que fossem acessados em uma outra etapa do processo de extração automatizado de dados.

Figura 15 – Processo de *web scrapping* do LinkedIn partindo do Google.



Fonte: elaborada pelo autor.

A primeira barreira encontrada foi o captcha do Google. O reCAPTCHA[23] usa um mecanismo avançado que visa impedir que *softwares* mal-intencionados se envolvam em atividades abusivas. Para tentar evitá-lo, utilizou-se o recurso de "*sleep*" do Python, que adiciona uma pausa no acesso automatizado ao site, de modo que simulasse uma pessoa real realizando buscas no Google. Ainda assim,

[22] Selenium. Disponível em: <https://www.selenium.dev>. Acesso em: 18 jul. 2024.
[23] reCAPTCHA. Disponível em: <https://www.google.com/recaptcha/about>. Acesso em: 18 jul. 2024.

durante a execução do *crawler* o reCAPTCHA aparecia, conforme mostrado na Figura 16.

Figura 16 – reCAPTCHA na busca do Google.

Fonte: Google, busca realizada pelo autor.

Caso a solução de extração de dados do LinkedIn partindo da busca do Google se mostrasse viável, para evitar bloqueios por parte do Google, muito provavelmente seria necessário utilizar o recurso de *proxy*, que é um serviço que atua como um intermediário entre o usuário e a Internet, recebendo e repassando as requisições ao site que está sendo acessado, ou seja, o *Internet Protocol* (IP) que é registrado nas páginas acessadas é o do *proxy* e não o da máquina do usuário.

O uso do Google foi descartado por alguns motivos, dentre eles, devido a quantidade de resultados mostrados como retorno da busca do Google (aproximadamente 28 mil), que não condizia com o número de perfis que realmente retornaram (30 páginas contendo 10 perfis cada, ou seja, 300 perfis). Vale ressaltar que outros testes foram feitos com o mecanismo de busca do Google, como por exemplo, passando vários nomes concatenados na expressão de busca (que futuramente poderiam vir de sistemas institucionais da UFSCar, como um portal de dados abertos), porém, o Google limita a expressão em 32 palavras.

A segunda tentativa de recuperar os perfis de ex-alunos da UFSCar no LinkedIn era por meio de busca na página UFSCar - Alumni.

O primeiro desafio enfrentado diz respeito a autenticação necessária para acessar informações no LinkedIn, que foi superado. Além disso, algumas tarefas

relacionadas ao procedimento de extração de dados já haviam sido desenvolvidas, como a busca de perfis de usuários e a seleção de metadados (também conhecida como “*parse*”) de um determinado perfil retornado.

Paralelamente a implementação técnica da solução, por meio de pesquisas na Internet, foi descoberto que o LinkedIn ganhou uma ação na justiça norte-americana contra a startup HiQ Labs[24], pois o uso de ferramentas automatizadas para extração de dados do LinkedIn vai contra os seus termos de uso[25] (item 8.2): “Desenvolver, dar suporte ou utilizar *software*, dispositivos, *scripts*, robôs ou quaisquer outros meios ou processos (incluindo *crawlers*, *plug-ins* e *add-ons* para navegadores ou outras tecnologias) para fazer varredura nos Serviços ou copiar de outra forma perfis e outros dados dos Serviços”. Sendo assim, optou-se por abortar a utilização dessa rede social como fonte de dados para análises sobre endogamia acadêmica.

A PL, por caracterizar-se como uma fonte de dados aberta, aliada à *expertise* do NIT-Materiais/UFSCar na extração e geração de indicadores dessa plataforma, mostrou-se ideal para a condução da pesquisa.

Para alcançar os objetivos geral e específicos desta pesquisa, o desenvolvimento do projeto foi estruturado em quatro macroatividades:

1. Construção do embasamento teórico;
2. Extração dos CVs Lattes e da produção científica de pesquisadores por meio das ferramentas csv_lattes e synclattes;
3. Utilização de *framework* Java[26] para serialização de objetos;
4. Aplicação de bibliotecas Python para geração de visualizações gráficas.

De acordo com Sant’Ana (2016), o acesso a dados tem crescido nos últimos anos em decorrência do aumento exponencial de soluções para coleta,

---

[24] HiQ Labs v. LinkedIn. Disponível em: <https://en.wikipedia.org/wiki/HiQ_Labs_v._LinkedIn>. Acesso em: 18 jul. 2024.
[25] Termos de Uso do LinkedIn. Disponível em: <https://br.linkedin.com/legal/user-agreement>. Acesso em: 18 jul. 2024.
[26] Jsonschema2pojo Core. Disponível em: <https://mvnrepository.com/artifact/org.jsonschema2pojo/jsonschema2pojo-core>. Acesso em: 10 nov. 2025.

armazenamento e recuperação de dados. Puerta-Díaz, Martí-Lahera e Martínez-Ávilla (2020) apontam que é comum que se tenha nas análises quantitativas dos subcampos dos estudos métricos da informação o uso de ferramentas para extração, tratamento e visualização de dados.

Nesta pesquisa foi utilizado o processo *Extract-Transform-Load* (ETL)[27], bastante conhecido na área de ciência de dados, que constitui as atividades de extração, transformação (tratamento) e carregamento de dados. É uma maneira amplamente utilizada para a combinação de dados de vários sistemas em um único banco de dados, armazenamento de dados ou *data lake*.

Sobre ciência de dados, para Provost e Fawcett (2013), é “um conjunto de princípios fundamentais que apoiam e orientam a extração de informações e conhecimento a partir de dados”. Foreman (2013) define ciência de dados como “a transformação de dados usando matemática e estatística em *insights*, decisões e produtos valiosos”. Ribeiro (2012), aponta que é muito importante a tarefa de transformar dados puros em informação e a partir disso facilitar a construção de conhecimento.

Esta pesquisa utilizou dois conjuntos distintos de dados. O primeiro conjunto de dados, que foi utilizado para testes de viabilidade técnica da solução, é proveniente da pesquisa de doutorado intitulada “Sistemática de Inteligência Acadêmica: caso das universidades públicas federais paulistas” (Sarvo, 2023), desenvolvida no âmbito do PPGCTS/UFSCar, sob orientação do Prof. Dr. Roniberto Morato do Amaral. A análise concentrou-se em registros de artigos publicados entre 2017 e 2020, obtidos a partir dos CVs Lattes de 3.077 docentes que atuam em PPGs nas universidades públicas federais paulistas.

O segundo conjunto de dados possui os CVs Lattes e a produção científica de 1.108 pesquisadores credenciados em PPGs dos campi da UFSCar em fevereiro de 2024, obtidos pelas ferramentas csv_lattes e synclattes, em uma extração (etapa “*Extract*” do ETL) realizada em novembro de 2025. Posteriormente, os currículos foram filtrados, sendo considerados para análises apenas os CVs Lattes dos pesquisadores da UFSCar – Campus São Carlos e suas produções científicas até o ano de 2024.

---

[27] ETL. Disponível em: <https://cloud.google.com/learn/what-is-etl>. Acesso em: 18 jul. 2024.

Vale ressaltar que para a extração dos CVs Lattes de forma automatizada, foi necessário possuir os IDs Lattes, os quais foram obtidos por meio *web scrapping* da Plataforma Sucupira. Tal recurso foi necessário, pois, até o momento, a CAPES não disponibiliza em seu portal de dados abertos[28] um conjunto de dados que contenha os IDs Lattes dos pesquisadores credenciados em PPGs.

Nesse contexto, segundo o Art. 1 da Resolução nº 510/2016 do Conselho Nacional de Saúde (CNS)[29], não serão registradas nem avaliadas pelo sistema CEP/Conep, itens como:

- Pesquisa que utilize informações de domínio público;
- Pesquisa com bancos de dados, cujas informações são agregadas, sem possibilidade de identificação individual.

Pelas especificidades desta pesquisa e considerando a Resolução CNS 510/2016, não foi necessário enviar a pesquisa para análise do Comitê de Ética da UFSCar.

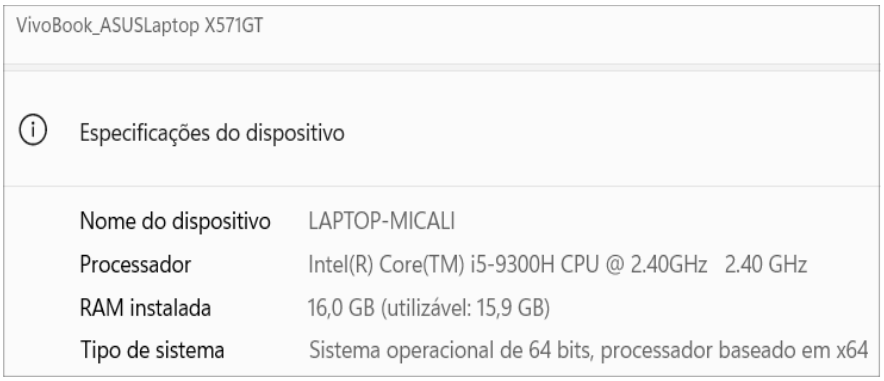
A respeito do armazenamento do conjunto de dados (etapa "*Load*" do ETL), foi analisada a possibilidade de trabalhar com NoSQL, que se refere a tipos não-relacionais de bancos de dados, os quais armazenam dados em um formato diferente das tabelas relacionais.

Compreende-se que *big data* possa ser um meio para realizar experimentação, pois as tecnologias utilizadas possuem como principal atrativo a capacidade de processar grandes quantidades de informações de forma ágil (Dumbill, 2012). Nesse sentido, a título de testes, foi realizada a importação dos dados de produção científica do primeiro conjunto de dados (50.968 artigos científicos em forma de documentos JSON), no banco de dados não-relacional MongoDB. O processo de importação durou em torno de 30 segundos. Para tanto, foi utilizado um computador com o sistema operacional Windows 11 e com a configuração de *hardware* descrita na Figura 17.

---

[28] Dados Abertos CAPES – Metadados de Docentes da Pós-Graduação. Disponível em: <https://dadosabertos.capes.gov.br/dataset/2021-a-2024-docentes-da-pos-graduacao-stricto-sensu-no-brasil/resource/ab445355-4a47-49be-995f-52c377f48225>. Acesso em: 20 jan. 2026.
[29] Resolução nº 510/2016 do CNS. Disponível em: <https://conselho.saude.gov.br/resolucoes/2016/Reso510.pdf>. Acesso em: 18 jul. 2024.

Figura 17 – Computador usado para a importação de dados no MongoDB.



Fonte: elaborada pelo autor.

O MongoDB, além de ser um banco de dados do tipo NoSQL, trabalha com arquivos flexíveis do tipo BSON (JSON binário) chamados de documentos, o que significa que os campos podem variar de documento para documento e a estrutura de dados pode ser alterada ao longo do tempo.

A nível de estudo de viabilidade técnica do uso do MongoDB, foram executados os seguintes procedimentos:

1. *Download* e instalação do MongoDB (serviço), MongoDB Shell (*prompt* de comandos) e MongoDB Compass (interface gráfica);
2. Tratamento dos dados extraídos pelas ferramentas csv_lattes e synclattes, via a IDE Sublime Text;
3. Importação dos CVs Lattes tratados, no formato JSON;
4. Criação de consultas para recuperação e manipulação de dados.

Optou-se por utilizar os dados no formato JSON, criado por Douglas Crockford[30] em 1999, baseado em JavaScript mas independente de linguagens e plataformas, sendo bastante utilizado em integrações entre sistemas, e, inclusive, pela API do ChatGPT[31].

Segundo Sant'Ana (2019), "uma vez armazenado o conteúdo, não se tem ainda garantia da presença das características que seriam necessárias para que

[30] Douglas Crockford. Disponível em: <http://www.crockford.com>. Acesso em: 18 jul. 2024.
[31] API do ChatGPT. Disponível em: <https://platform.openai.com/docs/guides/gpt/chat-completions-api>. Acesso em: 18 jul. 2024.

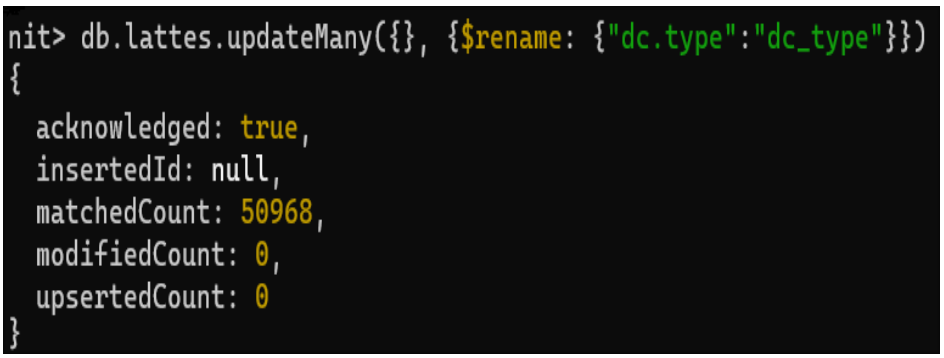
estes conteúdos sejam acessados no futuro da forma esperada". Sendo assim, a próxima etapa desta pesquisa envolveu o estudo de cada metadado que a ferramenta synclattes extraiu.

Em um primeiro momento, ao realizar consultas a nível de aprendizado no MongoDB Compass, percebeu-se que os dados não eram retornados. Após análise de cada um dos campos constantes no conjunto de dados, foi constatado que havia o caractere ponto (".") em alguns deles, e, a partir de pesquisas, descobriu-se que há uma limitação do MongoDB em relação à nome de campos contendo ponto.

Para ajudar no levantamento dos metadados utilizados pela ferramenta synclattes, foi consultada uma tabela que faz o "mapeamento dos campos do XML do Lattes utilizados para a composição dos campos Dublin Core Qualificado no DSpace" (Matias, 2015).

Após, foi feita uma tentativa de renomear os campos no MongoDB Shell, por exemplo, o campo "`dc.type`", como pode ser visto na Figura 18.

Figura 18 – Renomeando um campo pelo MongoDB Shell.



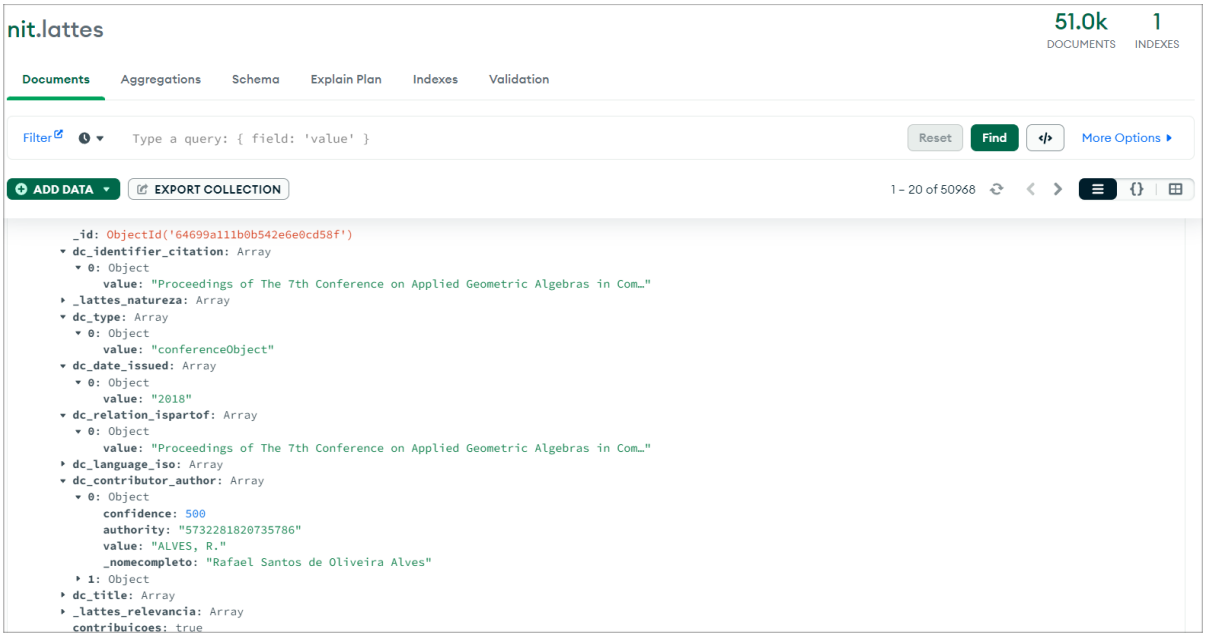
Fonte: elaborada pelo autor.

Como pode ser observado, nenhum registro foi renomeado com o uso do comando "`$rename`" (retornou "`modifiedCount: 0`"). Sendo assim, optou-se por fazer um tratamento nos dados por meio da IDE Sublime Text[32], com a opção "*replace whole word*", sendo substituídos, no nome dos campos, todos os caracteres "." por "_". Posteriormente, o conjunto de dados de produção científica tratado foi importado no MongoDB novamente.

[32] Sublime Text. Disponível em: <https://www.sublimetext.com>. Acesso em: 18 jul. 2024.

A Figura 19 apresenta uma tela do MongoDB Compass contendo o número aproximado de documentos importados (~51 mil), e, mais especificamente, os metadados de um desses documentos. Observa-se que os nomes dos campos não possuem mais o caractere ponto (“.”), por exemplo, o metadado “dc_identifier_citation”.

Figura 19 – CVs Lattes importados no MongoDB Compass.



Fonte: elaborada pelo autor.

Para confirmar que o problema com a questão do nome dos campos havia sido resolvido, após a importação dos dados tratados, foram executadas consultas no MongoDB Shell, conforme a Tabela 2.

Tabela 2: Consultas no MongoDB Shell.

Expressão	Data	Registros
`db.lattes.find({"_lattes_natureza": {$elemMatch: {value: "COMPLETO"}}}).count()`	16/08/2023	36.489
`db.lattes.find({"dc_contributor_author": {$elemMatch: {_nomecompleto: "Leandro Innocentini Lopes de Faria"}}}).count()`	16/08/2023	16

Fonte: elaborada pelo autor.

Indo ao encontro do que é argumentado por Cotta, Delbianco e Hilário (2020), que apontam que a indexação dos dados deve ser adequada para que seja possível a recuperação da informação de forma eficaz, a Figura 20 exibe um comando e o respectivo retorno após sua execução, que foi utilizado para adicionar um novo campo em um determinado documento JSON do conjunto de dados armazenado no MongoDB.

Figura 20 – Adicionando um novo campo em documento JSON no MongoDB.

Fonte: elaborada pelo autor.

Um exemplo de dado considerado nesta pesquisa para a interpretação de indicadores de endogamia é o endereço profissional que está disponível nos CVs Lattes, conforme é mostrado na Figura 21.

Figura 21 – Endereço profissional no CV Lattes.

Fonte: PL, acesso realizado pelo autor.

A PL incorpora uma validação do endereço profissional, que impede a associação de uma cidade a uma unidade federativa diferente daquela a que pertence, conforme pode ser vista na Figura 22.

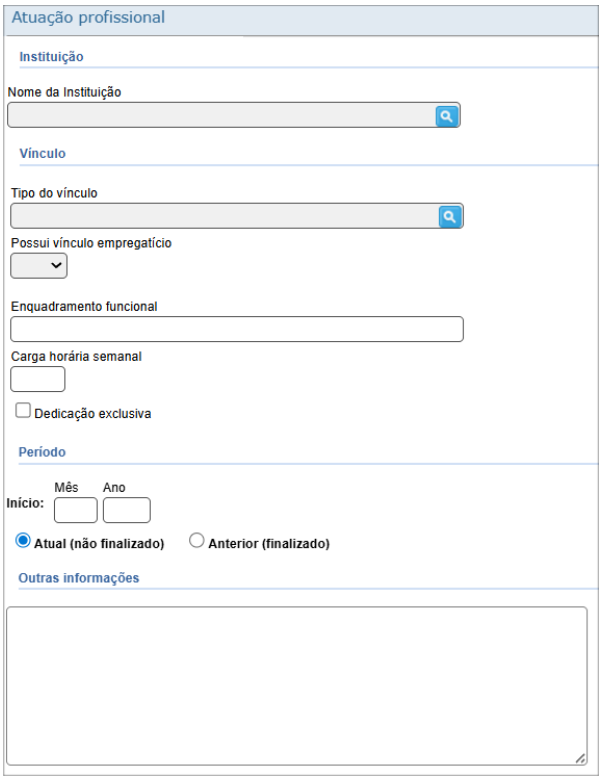
Figura 22 – Tela de validação do registro de endereço profissional na PL.

Fonte: PL, acesso realizado pelo autor.

A presença de mecanismos de validação, como a verificação de consistência entre cidade e unidade federativa no endereço profissional, representa uma vantagem significativa para análises baseadas nesses dados. Essa restrição, ao impedir a associação de municípios a estados incorretos, contribui para a integridade e a padronização geográfica das informações, reduzindo erros de origem humana e inconsistências cadastrais. Em contextos de pesquisas que envolvem o mapeamento territorial de instituições científicas e tecnológicas, essa característica garante maior confiabilidade às inferências espaciais, permitindo que a distribuição geográfica dos vínculos profissionais seja interpretada com precisão. Além disso, a validação facilita processos de geocodificação e integração com bases externas — como aquelas de natureza estatística ou administrativa —, uma vez que assegura a correspondência entre as unidades territoriais utilizadas. Dessa forma, o controle de consistência embutido na PL não apenas reflete uma racionalidade técnico-administrativa, mas também potencializa a qualidade analítica dos estudos que se fundamentam nos dados extraídos dessa base de dados.

Por outro lado, ainda no âmbito da PL, no formulário de cadastro de experiências profissionais falta um mecanismo de validação similar, conforme pode ser observado na Figura 23.

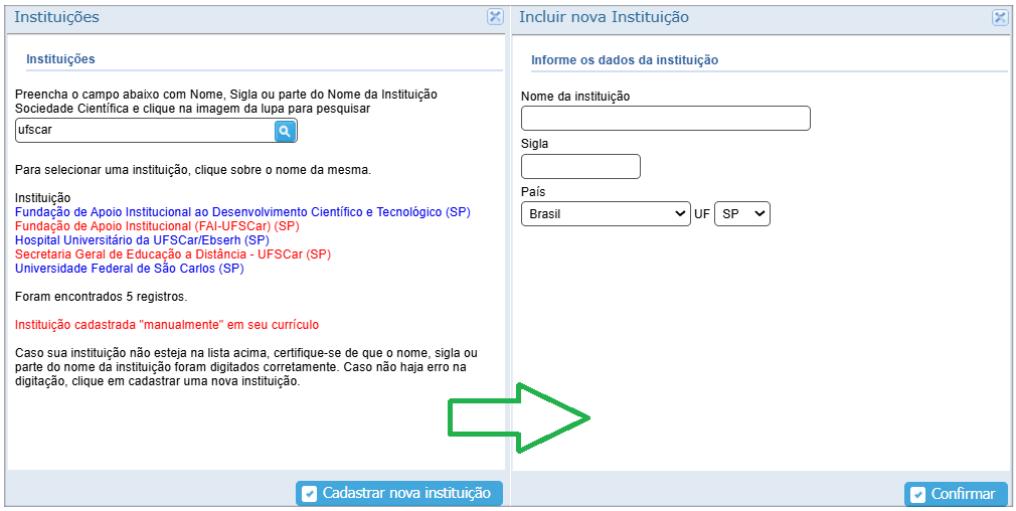
Figura 23 – Tela de cadastro de atuação profissional na PL.



Fonte: PL, acesso realizado pelo autor.

Ao clicar no ícone de busca (lupa) do campo nome da instituição, é aberta uma nova janela contendo uma tela de busca por instituições com um botão para cadastrar uma nova instituição, que, ao ser clicado, abre um formulário de inclusão de instituição, conforme pode ser visto na Figura 24.

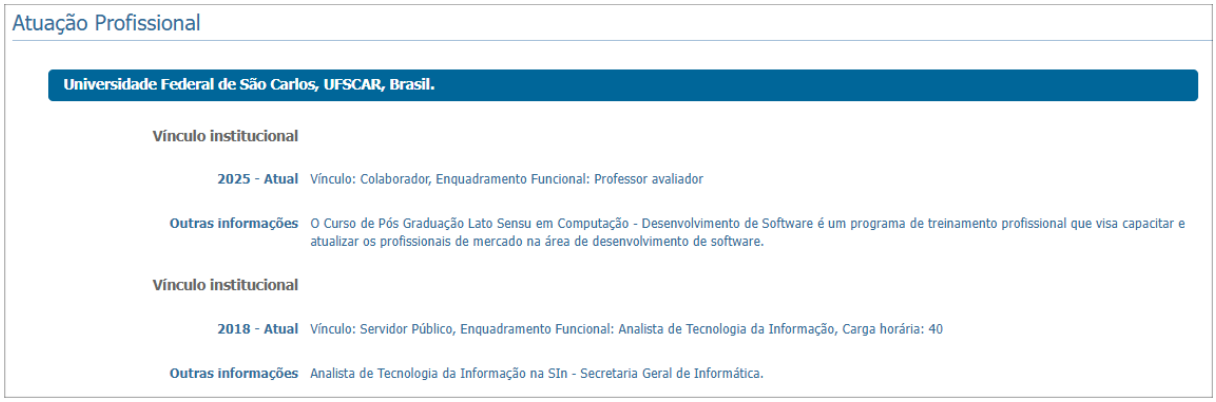
Figura 24 – Tela de cadastro de instituição na PL.



Fonte: PL, acesso realizado pelo autor.

Por fim, na Figura 25 há um trecho de um CV Lattes, onde consta um registro de atuação profissional.

Figura 25 – Registro de atuação profissional no CV Lattes.



Fonte: PL, acesso realizado pelo autor.

Conforme pode ser observado, o campo de cidade está ausente no formulário de cadastro de instituição e, consequentemente, nos registros de instituições da seção de atuação profissional dos CVs Lattes. Por essa razão, optou-se nessa pesquisa por trabalhar exclusivamente com o registro de endereço profissional do CV Lattes.

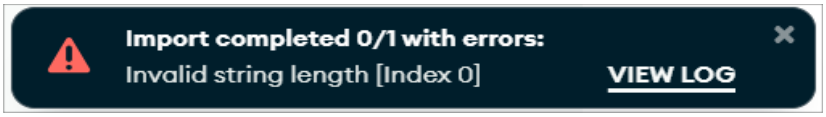
Durante o processo de extração de dados pela ferramenta synclattes, foi observado que o arquivo JSON de produção científica incorpora o identificador de unicidade (ID Lattes) dos coautores. Tal identificador atua como uma regra de autoridade de nomes, possibilitando a distinção precisa de pesquisadores e evitando ambiguidades decorrentes de homonímia ou variações de grafia. Essa característica é particularmente relevante em pesquisas sobre análise de redes de coautoria e mapeamento de colaborações científicas, uma vez que contribui para a integridade das relações entre autores e para a acurácia das inferências produzidas a partir dessas redes.

Após os testes com o primeiro conjunto de dados, iniciaram-se os estudos com o segundo conjunto de dados. Uma vez que o MongoDB trabalha com arquivos JSON e a ferramenta csv_lattes exporta os CVs Lattes no formato XML, foi necessário criar um *script* para converter os arquivos XML para JSON (Apêndice A). Ao realizar a importação dos CVs Lattes no MongoDB Compass, percebeu-se que a interface gráfica dele não permite a seleção de múltiplos arquivos. Para contornar

essa questão, foi implementado um *script* Python para a mesclagem de arquivos JSON (Apêndice B).

Ao tentar importar a produção científica dos pesquisadores da UFSCar no MongoDB, ocorreu um erro, que é exibido na Figura 26.

Figura 26 – Mensagem de erro ao importar dados no MongoDB.

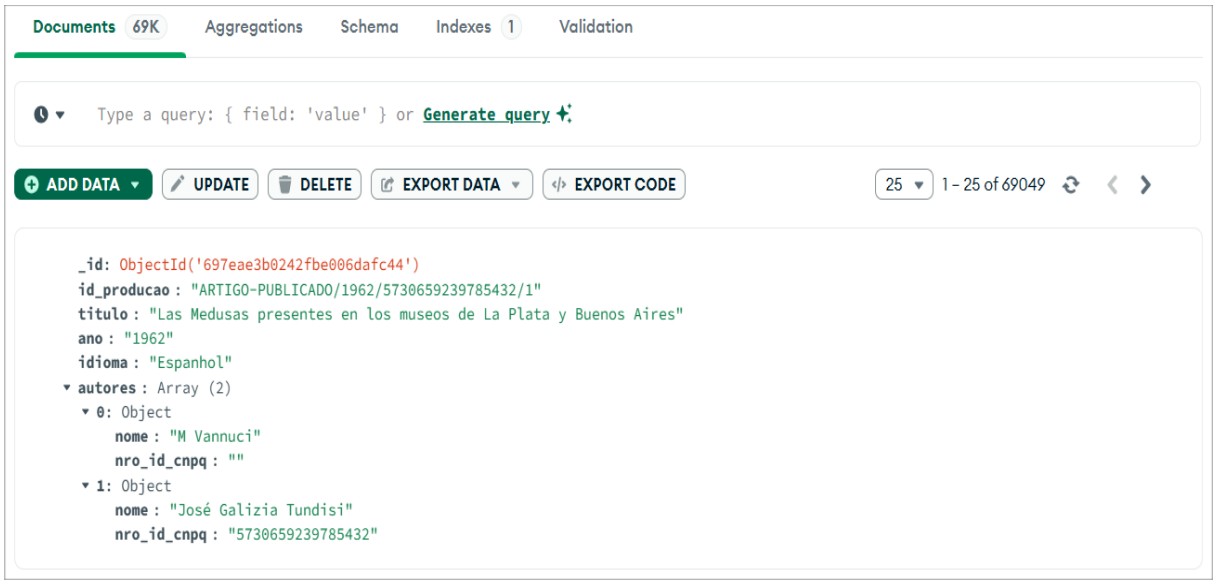


Fonte: elaborada pelo autor.

O log de erros retornou a mensagem "`Invalid string length [Index 0]`". Mediante análise, verificou-se que a estrutura do arquivo JSON de produção científica do segundo conjunto de dados era diferente daquela encontrada no primeiro conjunto de dados. Para contornar o problema, foi implementado um *script* Python (Apêndice C), que realiza o tratamento (etapa "*Transform*" do ETL) do conjunto de dados de produção científica.

A Figura 27 apresenta o resultado da importação desses dados no MongoDB, totalizando 69.049 documentos de artigos completos de revistas[33].

Figura 27 – Importação da produção científica da UFSCar no MongoDB.



Fonte: elaborada pelo autor.

[33] Dados tratados: Produção científica dos pesquisadores credenciados em PPGs da UFSCar. Disponível em: <https://zenodo.org/records/18446051>. Acesso em: 31 jan. 2026.

Apesar da viabilidade técnica de utilização do MongoDB, foram analisadas outras soluções, visando, principalmente, facilitar a geração de visualizações. Assim, optou-se por serializar os documentos JSON por meio de um *framework* Java.

A adoção de *framework* para serialização de objetos, em substituição ao uso direto de bancos de dados NoSQL como o MongoDB, representa uma estratégia tecnicamente vantajosa para projetos que envolvem o processamento e a análise de grandes volumes de dados estruturados em formato JSON. Essa abordagem é especialmente adequada em contextos de pesquisa e desenvolvimento de *software* em que se busca maior controle sobre as estruturas de dados, a modularidade da aplicação e a aderência a princípios da orientação a objetos. Os documentos JSON podem ser diretamente convertidos em classes Java, permitindo que o desenvolvedor trabalhe com tipos fortemente definidos, métodos encapsulados e hierarquias de herança. Essa modelagem favorece a consistência semântica dos dados e possibilita a aplicação de boas práticas de engenharia de *software*, como o reuso de código e a abstração de comportamentos.

Outro benefício relevante é a modularização e a separação de responsabilidades no código. Ao invés de concentrar a lógica de manipulação dos dados em consultas específicas de banco, cada parte do sistema pode ser responsável por uma função distinta do processo — leitura, tratamento, análise e persistência — o que torna a solução mais coesa e de manutenção mais simples. Essa estrutura modular também facilita a integração com outros componentes do procedimento computacional, inclusive com *scripts* acessórios desenvolvidos com outra linguagem de programação.

A extensibilidade e o reuso são igualmente fortalecidos. *Framework* de serialização permite a configuração de estratégias personalizadas para conversão de campos, formatação de datas, tratamento de exceções e filtragem de atributos, o que oferece flexibilidade para atender a diferentes cenários de uso. Além disso, os dados tratados em Java podem ser reutilizados em outros módulos ou projetos, contribuindo para a padronização e a interoperabilidade do código.

Em termos de desempenho, a abordagem baseada em processamento em memória proporciona ganhos expressivos de eficiência, uma vez que reduz a

dependência de operações de leitura em disco e permite o tratamento contínuo dos dados, sem interrupções causadas por gargalos de *Input*/*Output*.

Por fim, destaca-se a redução da dependência tecnológica e o consequente aumento da portabilidade do código. Ao substituir um sistema de banco de dados específico por um *framework* de serialização amplamente utilizado e de código aberto, o desenvolvedor assegura maior longevidade e reprodutibilidade à aplicação. Essa independência é especialmente relevante em contextos científicos, onde a transparência dos processos computacionais são requisitos fundamentais.

Em síntese, a utilização de um *framework* Java de serialização de objetos para manipulação de arquivos JSON constitui uma alternativa robusta e metodologicamente consistente, que alia desempenho, clareza arquitetural e aderência a princípios consolidados da engenharia de *software*.

O Apêndice D apresenta a classe para a serialização da produção científica e dos CVs Lattes. O primeiro teste de geração de classes a partir do arquivo JSON de CVs Lattes falhou, resultando no erro: "`[...] java.lang.OutOfMemoryError: Required array length [...] is too large`". Para contornar esse problema, foi necessário implementar um *script* Python (Apêndice E), para selecionar determinados campos dos CVs Lattes, uma vez que o *Framework* Java não conseguiu lidar com todos os metadados.

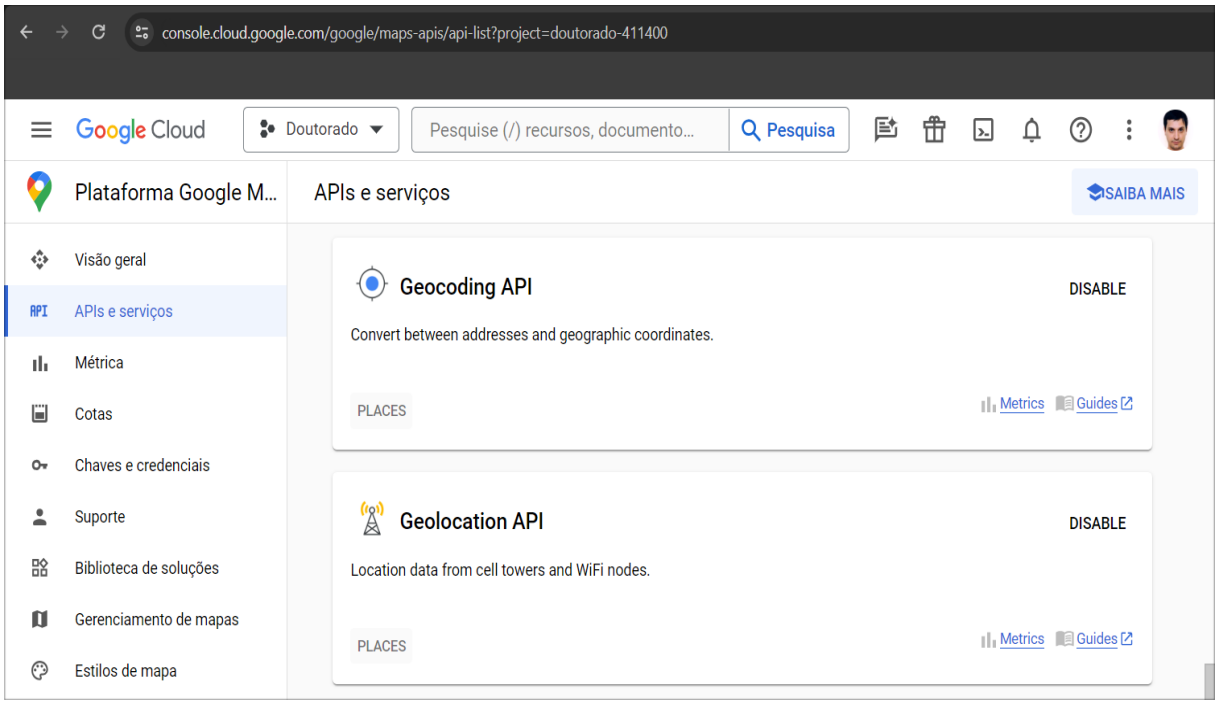
A etapa subsequente do procedimento computacional consiste na aplicação de recursos de geolocalização. Para tanto, adota-se o raio médio terrestre de 6.371 km (Smith, 2010) como constante de referência. Nesse contexto, a fórmula de Haversine é empregada para o cálculo de distâncias ortodrômicas entre coordenadas geográficas, sendo uma solução amplamente validada em sistemas de navegação e análise espacial (Johnson, 2015).

Contudo, essa abordagem apresenta limitações inerentes à simplificação do modelo. Ao assumir a Terra como uma esfera perfeita, a fórmula ignora a natureza de esferoide oblato do planeta, cujo raio varia em função da latitude (Pereira, 2012). Embora métodos mais complexos, como as fórmulas de Vincenty (1975), ofereçam maior precisão para distâncias extensas, a aproximação de Haversine satisfaz os requisitos desta pesquisa. O detalhamento do algoritmo implementado encontra-se disponível no Apêndice F.

No contexto de *big data*, com as devidas padronizações, é possível realizar a integração entre diferentes tecnologias, possibilitando a representação de dados de forma visual com a utilização de interfaces interativas, possuindo, entre outras características, filtros diversos. O Google, com a plataforma Google Cloud, fornece ferramentas modulares de alta tecnologia, possibilitando o uso de excelentes funcionalidades. A nível de validação de viabilidade para a presente pesquisa, uma conta grátis foi criada no Google Cloud Plataform e foi configurado um projeto chamado “Doutorado”, com dois serviços: “Geocoding API” e “Geolocation API”.

A Figura 28 exibe o console da plataforma Google Cloud e os serviços de geolocalização que foram habilitados.

Figura 28 – Console do Google Cloud Plataform.



Fonte: elaborada pelo autor.

Após a habilitação desses dois serviços, foi montada uma URL para testar se eles estavam ativos, conforme pode ser vista na Tabela 3.

Tabela 3: Expressão de teste de serviços do Google Cloud Plataform.

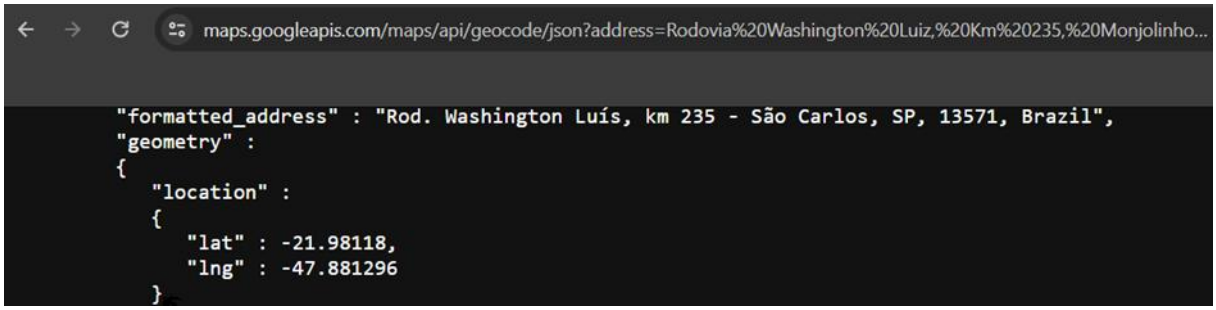
Expressão	Data	Reg.	Retorno
`https://maps.googleapis.com/maps/api/geocode/json?address=Rodovia Washington Luiz, Km 235, Monjolinho&key=`**`chave_privada`**	10/02/2024	1	JSON

Fonte: elaborada pelo autor.

Como pode ser observado, o último parâmetro da expressão que representa uma URL chama-se "*key*" e diz respeito a chave privada que a plataforma Google Cloud fornece para cada projeto criado (no caso, o projeto "Doutorado"). Para o parâmetro "*address*" foi passado o endereço da UFSCar.

A Figura 29 mostra uma parte do arquivo JSON que retornou ao executar a expressão de busca acima (URL) no navegador Google Chrome.

Figura 29 – Retorno do teste de serviços do Google Cloud Plataform.



Fonte: elaborada pelo autor.

Como pode ser observado, o arquivo JSON de retorno apresenta dados geográficos referentes ao endereço da UFSCar, como por ex., latitude ("lat") e longitude ("lng").

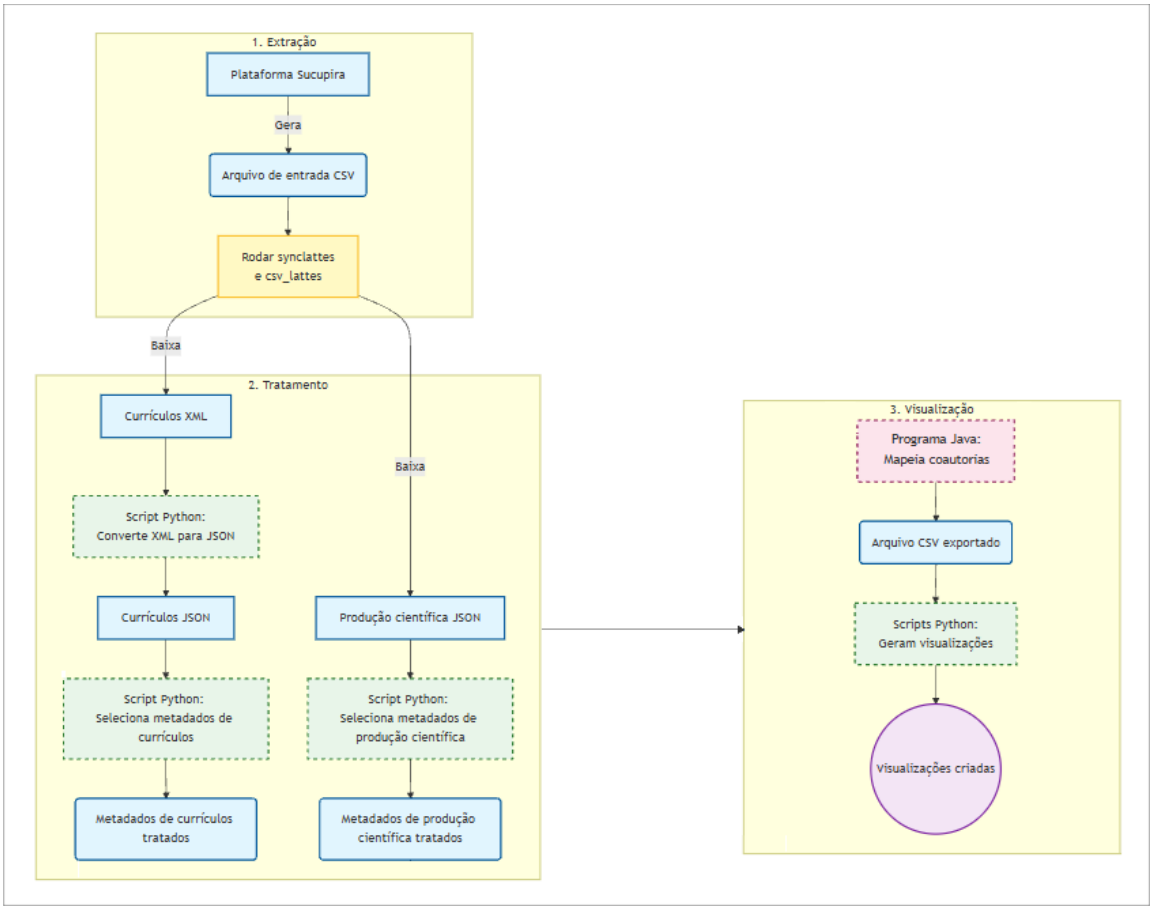
As APIs do Google Maps operam sob uma lógica comercial baseada em tarifação por volume de uso. Desde a reestruturação de seu modelo de preços em 2018, o Google passou a cobrar por requisições acima de determinados limites, impactando negativamente instituições públicas, educacionais e projetos de pesquisa que dependem de grandes volumes de dados geoespaciais (Google, 2025). Nesse sentido, essa política evidencia a assimetria entre plataformas corporativas e iniciativas de dados abertos, onde a dependência de soluções

proprietárias pode limitar a autonomia tecnológica e a sustentabilidade de iniciativas científicas e sociais.

Por sua vez, a API Nominatim, desenvolvida no âmbito do projeto OpenStreetMap, constitui uma ferramenta de geocodificação e busca espacial baseada em dados abertos, permitindo a conversão entre nomes de locais e coordenadas geográficas. Diferentemente de serviços proprietários, a Nominatim apoia-se em uma infraestrutura colaborativa e comunitária, refletindo os princípios do conhecimento livre e da ciência aberta. Sua utilização possibilita a incorporação de recursos de geolocalização em aplicações científicas, governamentais e educacionais, sem dependência de plataformas comerciais (Nominatim, 2025). Sendo assim, optou-se por utilizar a API Nominatim nesta pesquisa, conforme pode ser visto no Apêndice G.

Neste contexto, a Figura 30 apresenta o procedimento computacional desenvolvido nesta pesquisa.

Figura 30 – Procedimento de mapeamento de coautorias científicas.



Fonte: elaborada pelo autor.

O fluxo metodológico desenvolvido nesta pesquisa, conforme ilustrado, estrutura-se em três etapas sequenciais: Extração, Tratamento e Visualização. A etapa inicial, denominada Extração, obtém dados da Plataforma Sucupira e cria um arquivo de entrada[34] em formato CSV, contendo os IDs Lattes dos pesquisadores. Este arquivo alimenta as ferramentas csv_lattes e synclattes, responsáveis pela coleta massiva dos currículos[35] (em XML) e dos registros de produção científica[36] (em JSON).

Na etapa subsequente, Tratamento, um *script* desenvolvido em Python realiza a conversão dos currículos de XML para JSON (Apêndice A), garantindo a interoperabilidade dos dados. Outros *scripts* Python são então executados para a seleção dos metadados pertinentes de produção científica (Apêndice C) e de CVs Lattes (Apêndice E).

A etapa final, definida como Visualização, utiliza uma aplicação desenvolvida em linguagem Java para processar os dados tratados (Apêndice D), cruzar informações — mediante o campo ID Lattes, presente tanto no arquivo JSON de produção científica quanto no arquivo JSON de CVs Lattes — e mapear as publicações científicas (Apêndice H), identificando as relações de coautoria científica dos pesquisadores da UFSCar – Campus São Carlos credenciados em PPGs da universidade.

A lógica que foi definida estabelece que o pesquisador deve atender simultaneamente a dois critérios relacionados ao endereço institucional inserido no CV Lattes: (i) possuir vínculo com a UFSCar, identificado pelo código institucional “033500000006”, e (ii) trabalhar na cidade de São Carlos/SP. Apenas são considerados como pontos de origem das coautorias científicas analisadas os pesquisadores que satisfazem essas condições.

Para cada pesquisador elegível, são examinadas suas publicações científicas, previamente indexadas pelo ID Lattes, assegurando eficiência e precisão no acesso aos dados.

---

[34] IDs Lattes dos pesquisadores credenciados em PPGs da UFSCar. Disponível em: <https://zenodo.org/records/18371308>. Acesso em: 26 jan. 2026.
[35] CVs Lattes dos pesquisadores credenciados em PPGs da UFSCar. Disponível em: <https://zenodo.org/records/18371406>. Acesso em: 26 jan. 2026.
[36] Produção científica dos pesquisadores credenciados em PPGs da UFSCar. Disponível em: <https://zenodo.org/records/18371506>. Acesso em: 26 jan. 2026.

Conforme pode ser visto no Apêndice H, consideramos endogâmica uma coautoria de publicação científica (artigo completo de revista) que contiver:

Condição 1) autor orientador;

Condição 2) coautor orientado;

Condição 3) publicação após o doutorado do coautor orientado.

Importante ressaltar que foi considerado apenas o primeiro doutorado concluído pelo coautor, ou seja, em caso de ter feito mais de um doutorado, é considerado o mais antigo.

Do ponto de vista computacional, o processamento baseia-se na indexação prévia dos pesquisadores e das publicações em estruturas de dados do tipo Map, otimizando o desempenho das operações de busca. A análise geográfica emprega um mecanismo de *cache* em memória para evitar a recomputação de coordenadas geográficas, garantindo eficiência e escalabilidade. Cada relação de coautoria gera um registro contendo as coordenadas de origem e destino, bem como o ano da publicação, permitindo análises espaciais e temporais subsequentes.

Como etapa complementar ao processamento analítico, o sistema gera um arquivo auxiliar contendo exclusivamente IDs Lattes dos coautores cujos currículos ainda não se encontram disponíveis no conjunto de dados. Esse procedimento viabiliza uma estratégia iterativa e incremental de coleta de informações, fundamental para a ampliação progressiva da cobertura da rede científica analisada.

Durante o processamento das publicações, todos os coautores são inicialmente identificados. Em seguida, cada ID Lattes é confrontado com a base de currículos já carregada em memória. Os coautores ausentes no conjunto de dados são registrados em um arquivo CSV específico, contendo apenas uma coluna com IDs Lattes, para posterior obtenção dos CVs Lattes. A escrita desse arquivo ocorre de forma controlada, assegurando a unicidade dos identificadores, de modo que cada coautor seja listado uma única vez. Essa estratégia evita redundâncias, reduz o esforço computacional nas etapas subsequentes e previne a repetição desnecessária de consultas a PL. Do ponto de vista metodológico, esse mecanismo permite que o processo analítico seja conduzido de maneira progressiva: a cada nova iteração, os currículos coletados podem ser incorporados ao conjunto de

dados, refinando e expandindo o escopo da análise sem comprometer a coerência dos critérios previamente estabelecidos.

Por fim, o programa produz mais dois arquivos de saída[37]: um arquivo estruturado em formato CSV com coordenadas para a geração de visualizações cartográficas e um arquivo de log textual detalhado como mecanismo de rastreabilidade e transparência. Esses artefatos asseguram a auditabilidade, a reprodutibilidade e a confiabilidade do processo analítico.

A respeito do mapa de fluxos geográficos, podemos considerar que foi implementado em três etapas.

A primeira etapa, diz respeito ao processamento executado em Java (Apêndice H).

A segunda etapa (Apêndice I), utiliza o arquivo CSV contendo coordenadas geográficas de origem e destino, representadas pelas variáveis “latitude_origem”, “longitude_origem”, “latitude_destino” e “longitude_destino”. Esse arquivo é então processado no Python utilizando a biblioteca Pandas, sendo inicialmente carregado em uma estrutura de dados tabular (*Data Frame*). Em seguida, os registros são agrupados com base nas combinações únicas de coordenadas de origem e destino, permitindo identificar relações espaciais recorrentes no conjunto de dados. Para cada par origem–destino, é calculado o número de ocorrências, representando a frequência com que aquela relação geográfica se manifesta. O resultado desse agrupamento é exportado para um novo arquivo CSV, no qual cada linha corresponde a uma rota geográfica distinta, acompanhada do respectivo total de ocorrências. Importante ressaltar que, uma vez em posse desse arquivo, é possível utilizar qualquer tecnologia de InfoVis que lide com o formato CSV. No âmbito desta pesquisa, geraram-se dois arquivos CSV contendo as contagens de ocorrências geográficas: um para coautorias endogâmicas e outro para não endogâmicas[38].

Na terceira etapa, o arquivo de ocorrências é utilizado como insumo para a construção de um mapa interativo de fluxos geográficos. Para isso, foi desenvolvido

---

[37] Arquivos de saída da execução do programa de mapeamento de coautorias científicas dos pesquisadores da UFSCar – Campus São Carlos credenciados em PPGs da universidade. Disponível em: <https://zenodo.org/records/18446156>. Acesso em: 31 de janeiro de 2026.
[38] Arquivos CSV com contagens de ocorrências geográficas de coautorias endogâmicas e não endogâmicas. Disponível em: <https://zenodo.org/records/18446414>. Acesso em: 01 fev. 2026.

um *script* em Python (Apêndice J) com o uso das bibliotecas Pandas e Folium. Inicialmente, o arquivo de ocorrências é carregado e submetido a uma etapa de validação, na qual são descartados registros com valores ausentes nas coordenadas ou na quantidade de ocorrências, assegurando a consistência dos dados.

O mapa base é então inicializado a partir da média das coordenadas de origem, possibilitando um enquadramento espacial adequado da área analisada. Para cada relação origem–destino identificada, são inseridos marcadores pontuais representando os locais de origem e de destino, bem como uma linha conectando esses pontos, caracterizando visualmente o fluxo geográfico correspondente.

A espessura das linhas é definida de forma proporcional à quantidade de ocorrências associadas a cada rota, de modo que conexões mais frequentes sejam representadas por linhas visualmente mais destacadas. Além disso, setas direcionais são distribuídas ao longo das linhas, indicando o sentido do deslocamento entre origem e destino. As setas apresentam variação de cor e tamanho conforme faixas predefinidas de quantidade de ocorrências, reforçando a diferenciação visual entre fluxos de maior e menor intensidade.

Por fim, é incorporada ao mapa uma legenda explicativa, detalhando o significado das cores das setas e da espessura das linhas. O produto final consiste em um arquivo HTML interativo, que permite a exploração dinâmica das relações espaciais e das intensidades dos fluxos identificados.

Visando à caracterização estatística preliminar do conjunto de dados, foram desenvolvidos *scripts*, apresentados nos Apêndices K a N, que exploram a incidência da endogamia em diferentes níveis de agregação. A análise inicia-se no nível mais granular com o Apêndice K, que quantifica a proporção global de relações de coautoria classificadas como endogâmicas e não endogâmicas. Elevando o nível de análise, o Apêndice L categoriza as publicações entre aquelas que apresentam incidência de endogamia e as que não possuem qualquer coautoria endogâmica. Para refinar essa métrica, o Apêndice M abandona a classificação binária e apresenta, por meio de histogramas, a distribuição de frequência das publicações conforme sua taxa percentual de endogamia. Por fim, focando nos pesquisadores, o Apêndice N mostra a frequência com que eles estabelecem parcerias endogâmicas.

Para operacionalizar a análise visual dos dados, foram criados *scripts*, exibidos nos Apêndices O a S. A investigação tem início pelo comportamento institucional no Apêndice O, que examina a relação entre a quantidade de publicações e a porcentagem de coautorias endogâmicas. Em seguida, o Apêndice P introduz a variável geográfica — adotando a UFSCar – Campus São Carlos como referência de origem — para analisar a influência da distância física sobre a porcentagem de coautorias endogâmicas das instituições. O Apêndice Q amplia o escopo espacial para o volume absoluto, mapeando a distribuição da quantidade total de coautorias das instituições em função da distância da UFSCar – Campus São Carlos. Por fim, para qualificar essa dispersão territorial, o Apêndice R isola o volume de publicações especificamente endogâmicas, permitindo o contraste direto com o Apêndice S, que apresenta a distribuição geográfica das publicações não endogâmicas.

Com o objetivo de analisar a relação entre o nível de endogamia e a distância geográfica média entre os pesquisadores da UFSCar – Campus São Carlos credenciados em PPGs da universidade e seus coautores, foi desenvolvido um *script* em linguagem Python, apresentado no Apêndice T.

Na primeira etapa do *script*, os dados do arquivo CSV com coordenadas são carregados em memória utilizando a biblioteca Pandas. Posteriormente, a classificação categórica de endogamia é convertida em uma variável numérica auxiliar, denominada “*is_endogamica*”. Nessa conversão, as publicações classificadas como endogâmicas recebem valor 1, enquanto as não endogâmicas recebem valor 0. Essa transformação permite o cálculo do nível de endogamia por meio de operações estatísticas simples, em especial a média aritmética.

Na sequência, os dados são agregados por produção científica, a partir do agrupamento pelo identificador “id_producao”. Para cada publicação, são calculados dois indicadores principais: (i) o nível médio de endogamia, obtido pela média dos valores da variável “*is_endogamica*”, e (ii) a distância geográfica média entre os autores e coautores, calculada a partir da média da variável “distancia_km”. O nível médio de endogamia varia no intervalo contínuo entre 0 e 1, em que valores próximos de 0 indicam ausência ou baixa endogamia, enquanto valores próximos de 1 indicam alta concentração de coautorias endogâmicas.

O *script* produz um gráfico de dispersão que representa visualmente a relação entre o nível de endogamia e a distância geográfica média. Nesse gráfico, cada ponto corresponde a uma publicação científica, sob a perspectiva de um determinado pesquisador da UFSCar – Campus São Carlos. A incorporação da dimensão temporal através do filtro de seleção do ano, transformou a visualização em um instrumento capaz de identificar tendências históricas, deixando de ser apenas um registro acumulado da produção científica para se tornar um meio de diagnóstico, permitindo observar se o comportamento da rede se altera ao longo dos anos. Essa funcionalidade é essencial para distinguir se determinados agrupamentos de alta endogamia são características estruturais permanentes da instituição ou fenômenos isolados em períodos específicos, oferecendo uma profundidade analítica que tabelas ou gráficos estáticos não conseguem prover.

4 RESULTADOS E DISCUSSÃO

Conforme elucidado, o programa de mapeamento de coautorias científicas exporta três arquivos, sendo um deles um arquivo com IDs Lattes dos coautores. Esse arquivo[39], gerado pelo primeiro processamento desse programa (Apêndice H) e executado sem a limitação temporal definida pelo parâmetro “ANO_LIMITE” (adicionado posteriormente), contém os IDs Lattes de 8.293 coautores dos pesquisadores da UFSCar – Campus São Carlos credenciados em PPGs da universidade. Com base nessa listagem, foi realizada em janeiro de 2026 uma nova coleta de dados utilizando a ferramenta csv_lattes. Visto que não era necessária a extração da produção científica desses coautores nessa etapa (dispensando o uso da synclattes), a extração resultou em 8.250 CVs Lattes válidos[40].

Com os CVs Lattes dos coautores extraídos, repetiu-se o procedimento de mapeamento de coautorias científicas, com apenas uma diferença: após a conversão desses currículos — do formato XML para o formato JSON (Apêndice A) — e a extração de metadados (Apêndice E – variável “credenciado” com o valor ‘N’), foi necessário mesclar (Apêndice B) os CVs Lattes dos pesquisadores credenciados em PPGs da UFSCar com os CVs Lattes dos seus coautores, ambos já tratados[41]. Em seguida, a execução do programa de mapeamento de coautorias (Apêndice H) processou 86.341 registros de coautorias em artigos completos de revistas até o ano de 2024, referentes a 24.612 produções únicas. Esses dados englobam 664 pesquisadores da UFSCar – Campus São Carlos credenciados em PPGs da universidade e 7.322 coautores com endereço profissional cadastrado.

Em relação ao *script* de geração de mapa interativo (Apêndice J), a Figura 31 exibe um mapa com os fluxos de coautoria científica endogâmica dos pesquisadores da UFSCar – Campus São Carlos credenciados em PPGs da universidade.

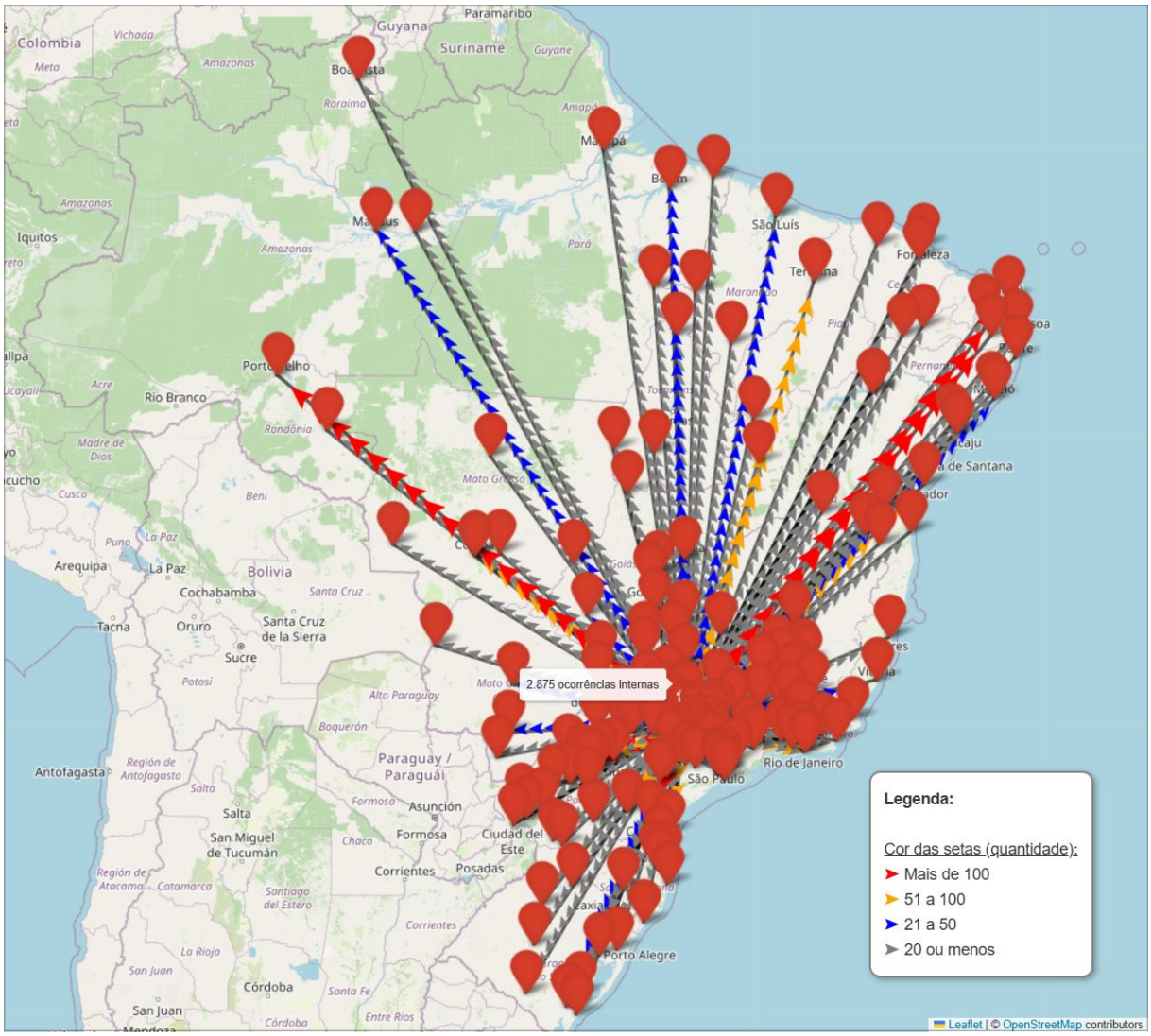
---

[39] IDs Lattes dos coautores dos pesquisadores da UFSCar – Campus São Carlos credenciados em PPGs da universidade. Disponível em: <https://zenodo.org/records/18371357>. Acesso em: 26 jan. 2026.

[40] CVs Lattes dos coautores dos pesquisadores da UFSCar – Campus São Carlos credenciados em PPGs da universidade. Disponível em: <https://zenodo.org/records/18371450>. Acesso em: 26 jan. 2026.

[41] Dados tratados: CVs Lattes dos pesquisadores credenciados em PPGs da UFSCar e dos coautores dos pesquisadores da UFSCar – Campus São Carlos credenciados em PPGs da universidade. Disponível em: <https://zenodo.org/records/18446038>. Acesso em: 31 jan. 2026.

Figura 31 – Distribuição espacial dos fluxos de coautoria científica endogâmica dos pesquisadores credenciados em PPGs da UFSCar.

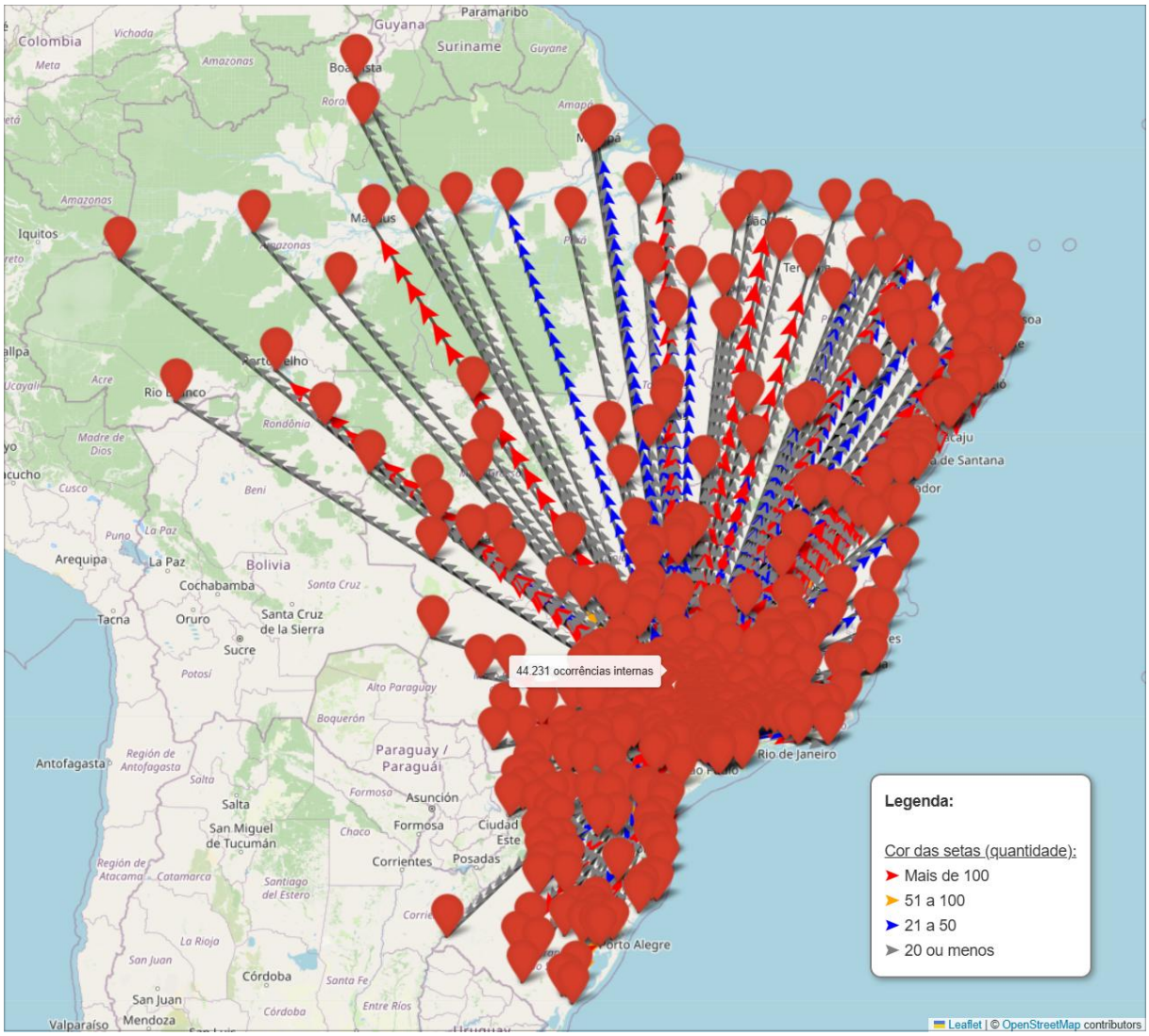


Fonte: elaborada pelo autor.

Para evitar o ofuscamento das conexões externas, optou-se por padronizar a espessura das arestas, fornecendo a informação do volume interno por meio de um recurso de *tooltip* (exibido ao posicionar o *cursor* sobre o ícone do campus, representado pelo ícone *home*). Essa estratégia preserva a clareza dos fluxos interinstitucionais, mantendo acessível a informação de coautoria local. Como pode ser observado, há 2.875 coautorias endogâmicas internas, ou seja, tendo a UFSCar – Campus São Carlos como origem e destino.

Na Figura 32 é apresentado um mapa com os fluxos de coautoria científica não endogâmica dos pesquisadores da UFSCar – Campus São Carlos credenciados em PPGs da universidade.

Figura 32 – Distribuição espacial dos fluxos de coautoria científica não endogâmica dos pesquisadores credenciados em PPGs da UFSCar.

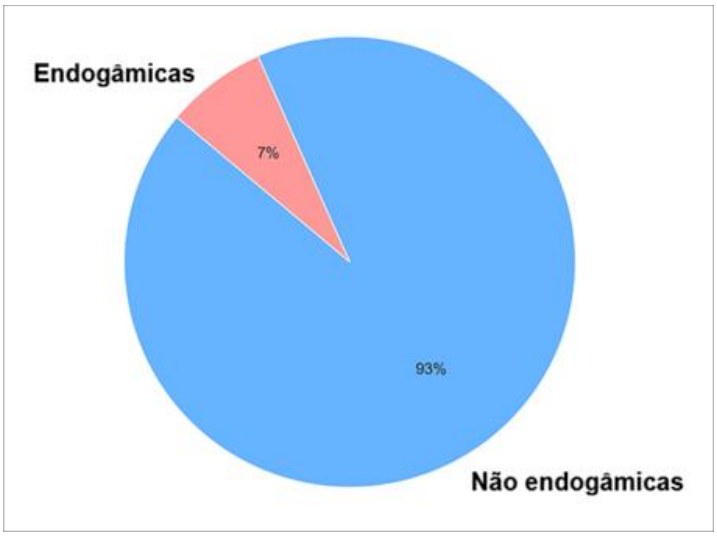


Fonte: elaborada pelo autor.

Conforme pode ser visto, há mais registros de coautorias não endogâmicas, porém a topologia de rede é semelhante.

A seguir, são apresentadas as visualizações referentes aos apêndices K a S.

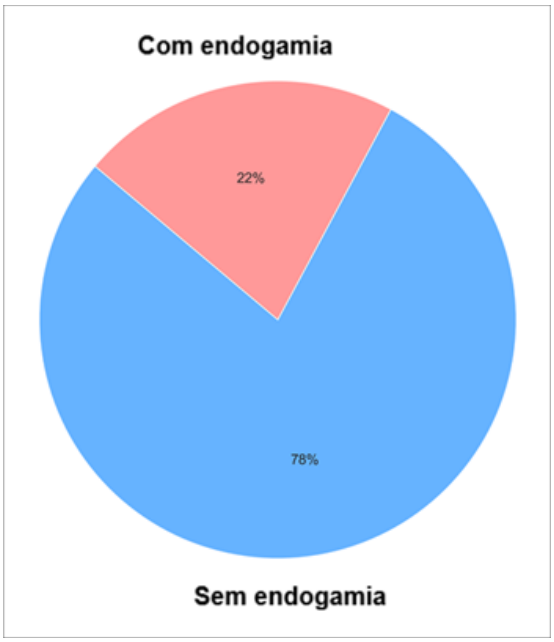
Figura 33 – Visualização da porcentagem de endogamia das coautorias.



Fonte: elaborada pelo autor.

A Figura 33 mostra a distribuição das coautorias classificadas como endogâmicas e não endogâmicas no conjunto de dados analisado. Observa-se o predomínio expressivo de coautorias não endogâmicas, que correspondem a 93% do universo investigado, enquanto as coautorias endogâmicas representam apenas 7%, indicando que a endogamia se configura como um fenômeno minoritário no conjunto das coautorias analisadas.

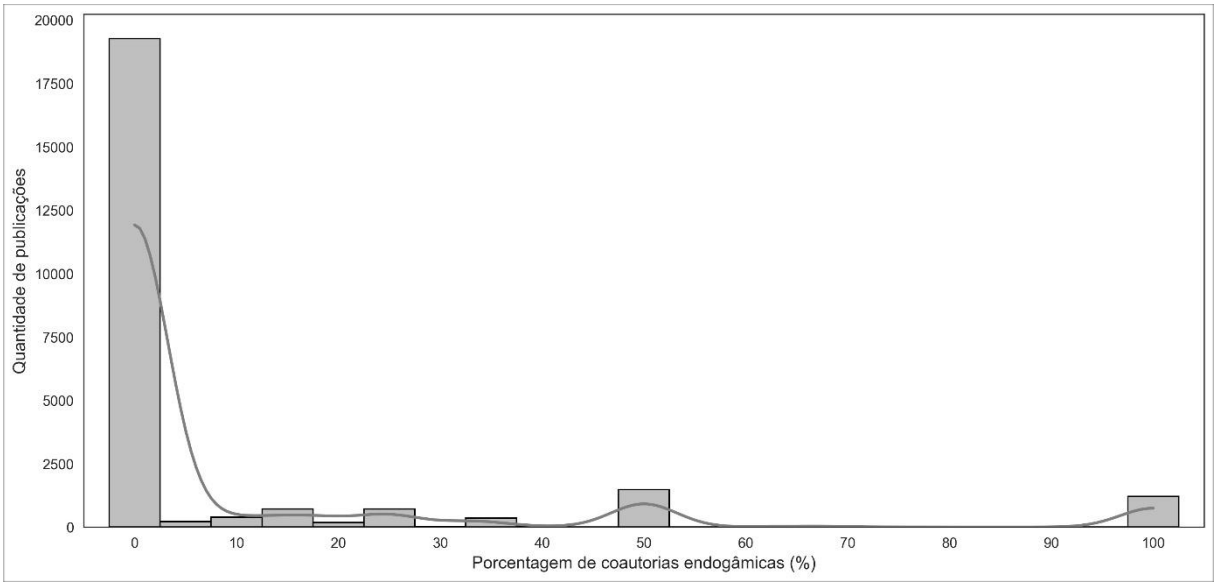
Figura 34 – Visualização da porcentagem de endogamia das publicações.



Fonte: elaborada pelo autor.

A Figura 34 exibe a proporção de publicações com e sem endogamia no conjunto de dados analisado. Verifica-se que 22% das publicações apresentam endogamia, enquanto 78% não a manifestam, evidenciando que, embora minoritária, a endogamia assume maior expressividade quando observada a partir da unidade de análise das publicações.

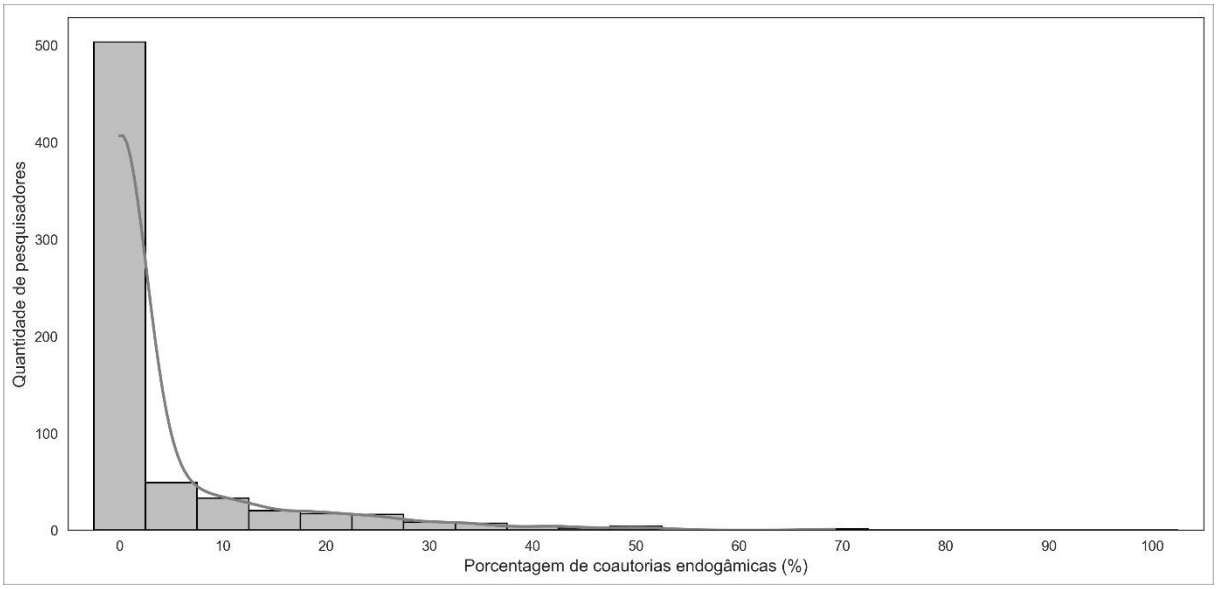
Figura 35 – Visualização da endogamia das publicações.



Fonte: elaborada pelo autor.

A Figura 35 revela a distribuição da endogamia das publicações, expressa em termos percentuais. Observa-se uma forte concentração de publicações nos valores mais baixos de endogamia, com predominância próxima a 0%, indicando que a maioria das publicações não apresenta vínculos endogâmicos ou os manifesta de forma muito reduzida. Ao mesmo tempo, identifica-se a presença de picos isolados em faixas mais elevadas — incluindo valores intermediários e casos extremos próximos a 100% —, o que evidencia a existência de publicações com elevados níveis de endogamia. Essa distribuição assimétrica reforça que, embora a endogamia seja minoritária no conjunto das publicações, ela se manifesta de forma concentrada e intensa em casos específicos, mostrando heteregoneidade no comportamento das coautorias.

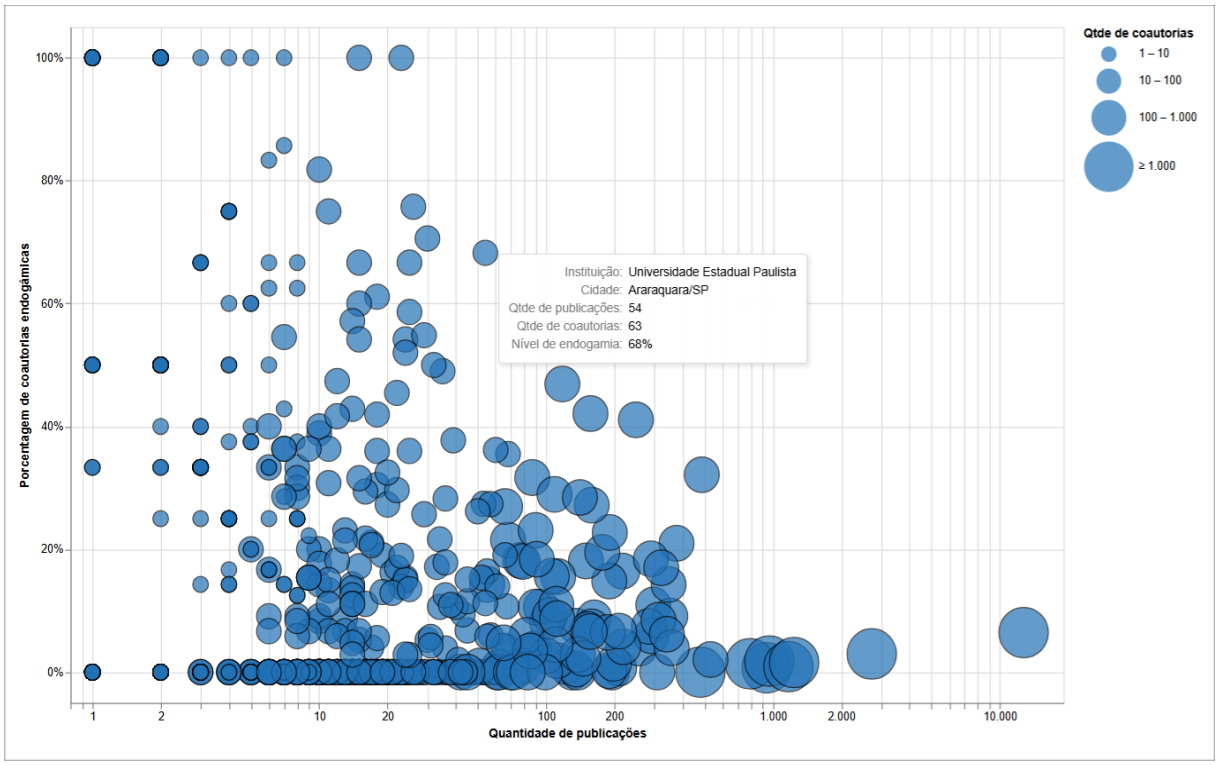
Figura 36 – Visualização da endogamia dos pesquisadores.



Fonte: elaborada pelo autor.

A Figura 36 apresenta a distribuição da endogamia dos pesquisadores, expressa pela porcentagem de coautorias endogâmicas associadas a cada indivíduo. Observa-se uma concentração acentuada de pesquisadores nas faixas mais baixas de endogamia, especialmente próximas a 0%, indicando que a maioria dos pesquisadores estabelece predominantemente coautorias não endogâmicas. À medida que a porcentagem de endogamia aumenta, verifica-se uma redução progressiva na quantidade de pesquisadores, com poucos casos distribuídos em faixas mais elevadas. Essa distribuição assimétrica evidencia que, ainda que a endogamia não seja dominante no comportamento individual dos pesquisadores, há casos específicos em que ela se manifesta de forma mais intensa, reforçando a heterogeneidade das dinâmicas de coautoria no conjunto de dados analisado.

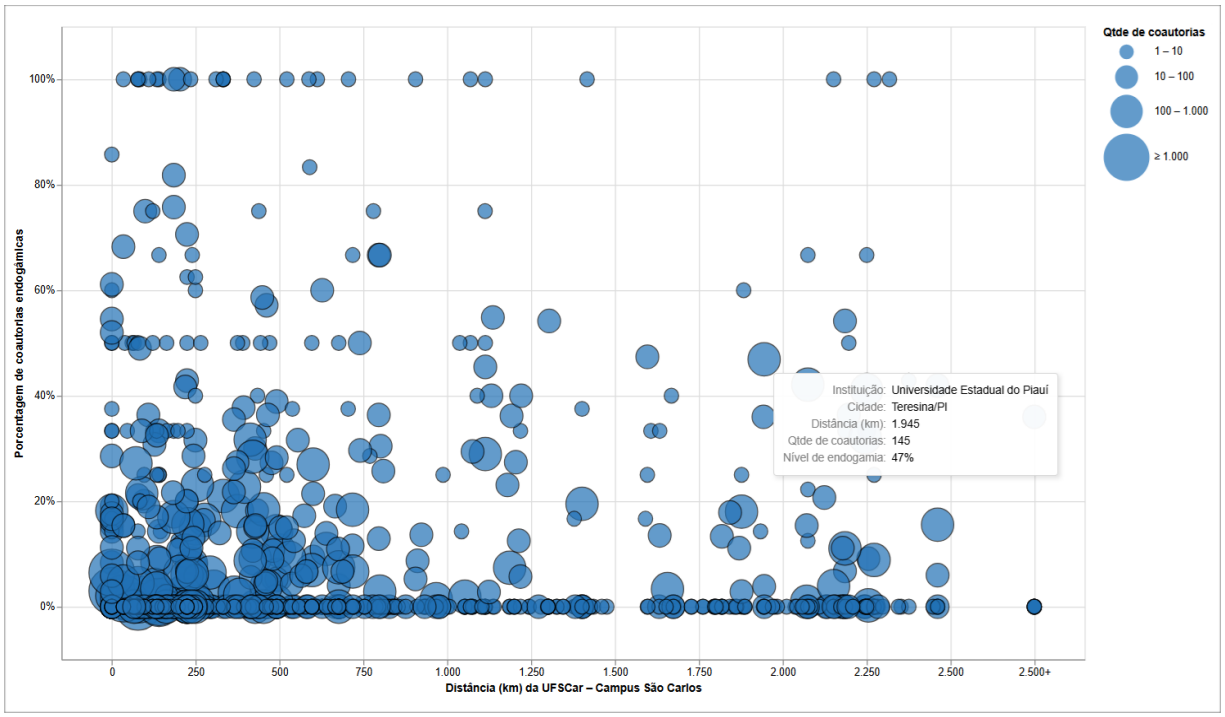
Figura 37 – Visualização da endogamia das instituições.



Fonte: elaborada pelo autor.

A Figura 37 exibe a distribuição da endogamia das instituições, relacionando a quantidade de publicações ao percentual de coautorias endogâmicas, com o tamanho dos marcadores indicando o volume de coautorias. Observa-se que instituições com maior produção científica tendem a concentrar-se nas faixas mais baixas de endogamia, especialmente próximas a 0%, indicando maior abertura e diversificação das redes de colaboração. Em contraste, níveis mais elevados de endogamia aparecem de forma mais frequente entre instituições com menor volume de publicações, embora existam casos pontuais de instituições com produção intermediária ou elevada que apresentam percentuais significativos de endogamia. Esse indicador sugere uma relação inversa entre volume de produção e intensidade da endogamia institucional, ao mesmo tempo em que evidencia a heterogeneidade dos tipos de coautoria.

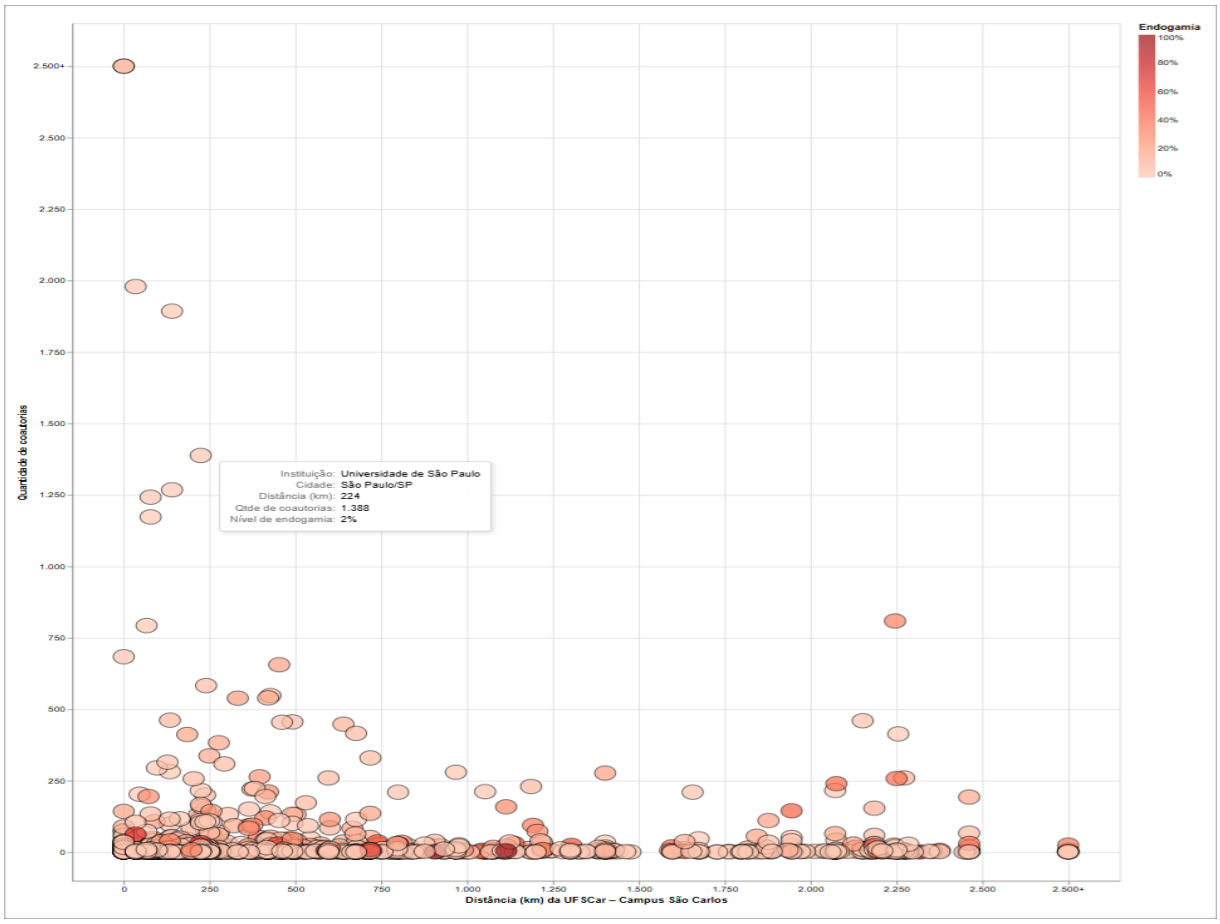
Figura 38 – Visualização da endogamia das instituições por distância.



Fonte: elaborada pelo autor.

A Figura 38 apresenta a distribuição da endogamia das instituições em função da distância geográfica em relação à UFSCar – Campus São Carlos, relacionando a porcentagem de coautorias endogâmicas à distância (em km), com o tamanho dos marcadores indicando o volume de coautorias. Observa-se uma concentração expressiva de instituições localizadas a menores distâncias nas faixas mais baixas de endogamia, especialmente próximas a 0%, sugerindo que a proximidade geográfica, por si só, não implica necessariamente maiores níveis de endogamia. Ao longo de toda a extensão do eixo de distância, inclusive em instituições mais distantes, identificam-se casos com diferentes níveis de endogamia, sem a configuração de uma característica linear evidente. Esse resultado indica que a endogamia institucional não é determinada exclusivamente pela proximidade espacial, mas resulta de dinâmicas mais complexas de colaboração científica.

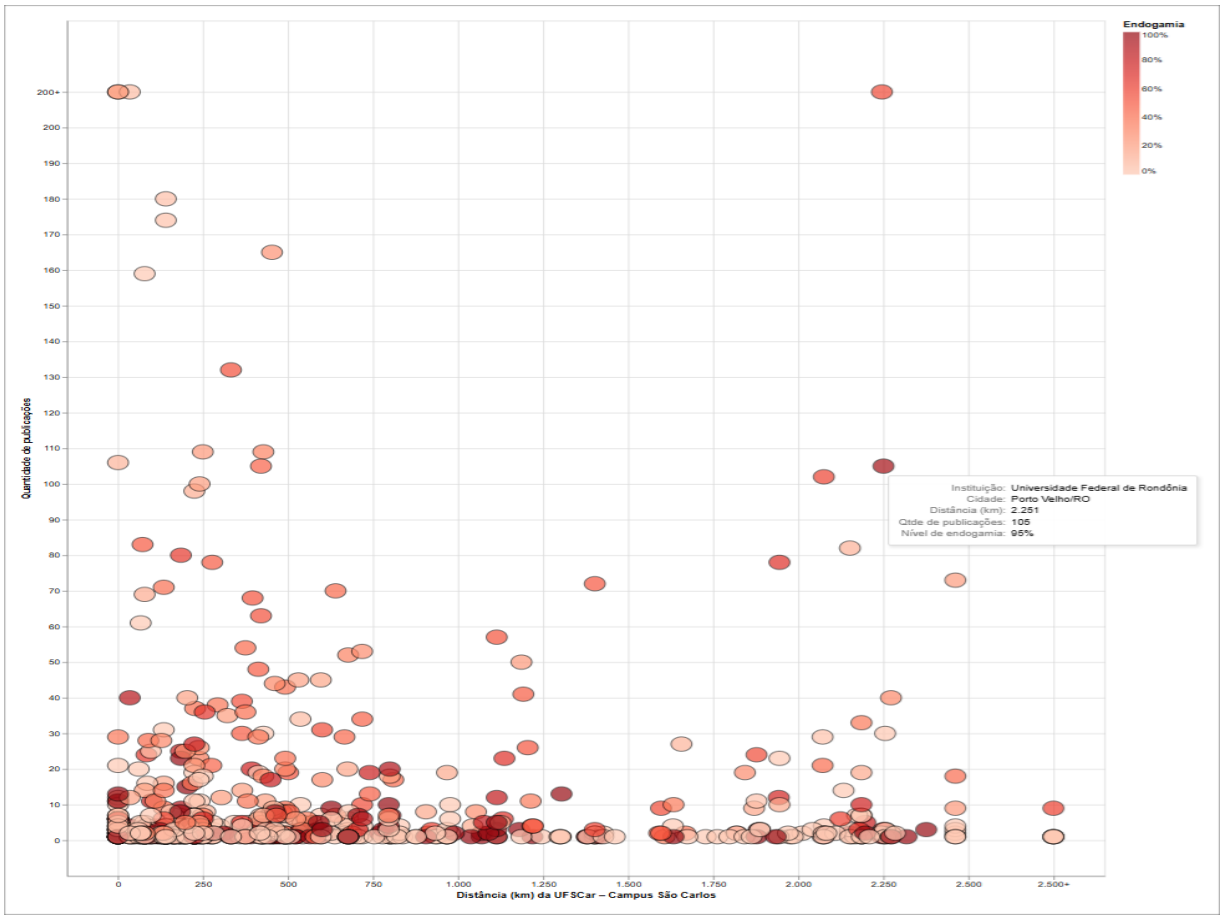
Figura 39 – Visualização da coautoria das instituições.



Fonte: elaborada pelo autor.

A Figura 39 exibe a distribuição das coautorias institucionais em função da distância geográfica em relação à UFSCar – Campus São Carlos, relacionando a quantidade de coautorias à distância (em km), com a coloração dos marcadores indicando o nível de endogamia. Observa-se que as maiores quantidades de coautorias se concentram entre instituições localizadas a menores distâncias, especialmente nos primeiros intervalos do eixo espacial, ainda que essas instituições apresentem, em geral, baixos níveis de endogamia. À medida que a distância aumenta, verifica-se uma redução na intensidade das coautorias, acompanhada por uma distribuição mais dispersa dos níveis de endogamia. Essa característica reforça que a proximidade geográfica favorece o volume de colaboração institucional, mas não implica, necessariamente, maior endogamia, corroborando os resultados observados nas figuras anteriores.

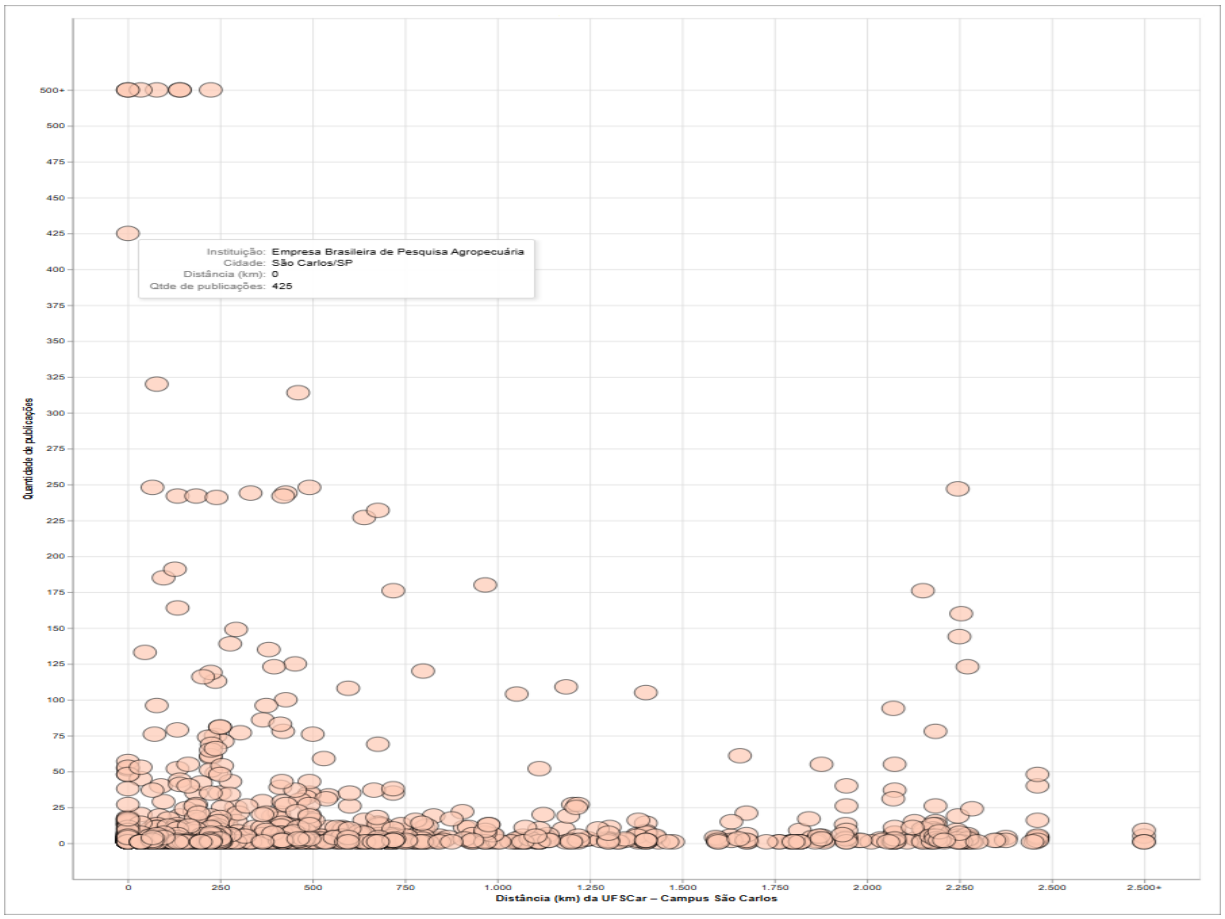
Figura 40 – Visualização do total de publicações endogâmicas das instituições.



Fonte: elaborada pelo autor.

A Figura 40 apresenta a distribuição das instituições segundo a quantidade total de publicações endogâmicas em função da distância geográfica em relação à UFSCar – Campus São Carlos, com a coloração dos marcadores indicando o nível de endogamia institucional. Observa-se que a maior parte das instituições se concentra em faixas reduzidas de produção, independentemente da distância, enquanto poucas instituições apresentam volumes mais elevados de publicações, distribuídas tanto em menores quanto em maiores distâncias. A variação cromática evidencia que níveis distintos de endogamia estão presentes ao longo de toda a distribuição, sem um relação clara entre quantidade de publicações, distância geográfica e intensidade da endogamia. Esse resultado sugere que o volume de produção institucional, isoladamente, não determina os níveis de endogamia observados.

Figura 41 – Visualização do total de publicações não endogâmicas das instituições.



Fonte: elaborada pelo autor.

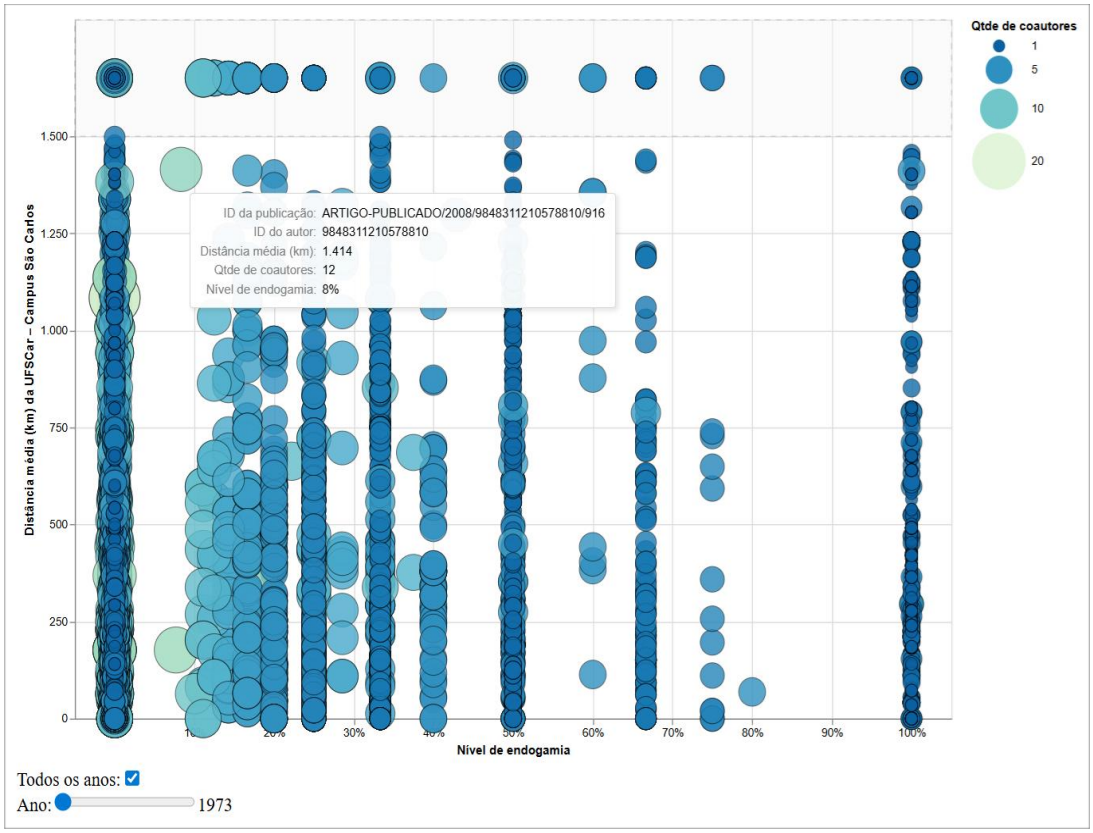
A Figura 41 mostra a distribuição das instituições segundo a quantidade total de publicações não endogâmicas em função da distância geográfica em relação à UFSCar – Campus São Carlos. Observa-se que a distribuição das publicações não endogâmicas não apresenta associação direta com a distância geográfica, indicando que esse tipo de produção se distribui de forma heterogênea ao longo de todo o eixo espacial considerado.

De forma sintética, as análises apresentadas nas Figuras 33 a 41 evidenciam que a endogamia nas coautorias e publicações constitui um fenômeno minoritário no conjunto de dados, embora se manifeste de maneira concentrada em casos específicos, variando conforme a unidade de análise adotada. Observa-se que a endogamia tende a ser mais expressiva quando examinada em recortes institucionais ou individuais, ao passo que, no plano agregado, predomina a colaboração não endogâmica. Ademais, os resultados indicam que não há uma associação direta e linear entre endogamia, distância geográfica e volume de

produção científica, sugerindo que as dinâmicas de colaboração são condicionadas por fatores institucionais, estruturais e relacionais mais complexos. Esse conjunto de evidências fornece o pano de fundo analítico para a interpretação das Figuras 42 e 43, que aprofundam a compreensão dessas dinâmicas sob a perspectiva dos pesquisadores.

Na Figura 42 é exibida uma visualização gráfica dinâmica para análise do nível de endogamia e distância média dos pesquisadores, gerada pelo *script* Python do Apêndice T, que, por meio do arquivo CSV de coordenadas, manipulou publicações do ano de 1973 a 2024.

Figura 42 – Visualização dinâmica do nível de endogamia e distância geográfica dos pesquisadores.



Fonte: elaborada pelo autor.

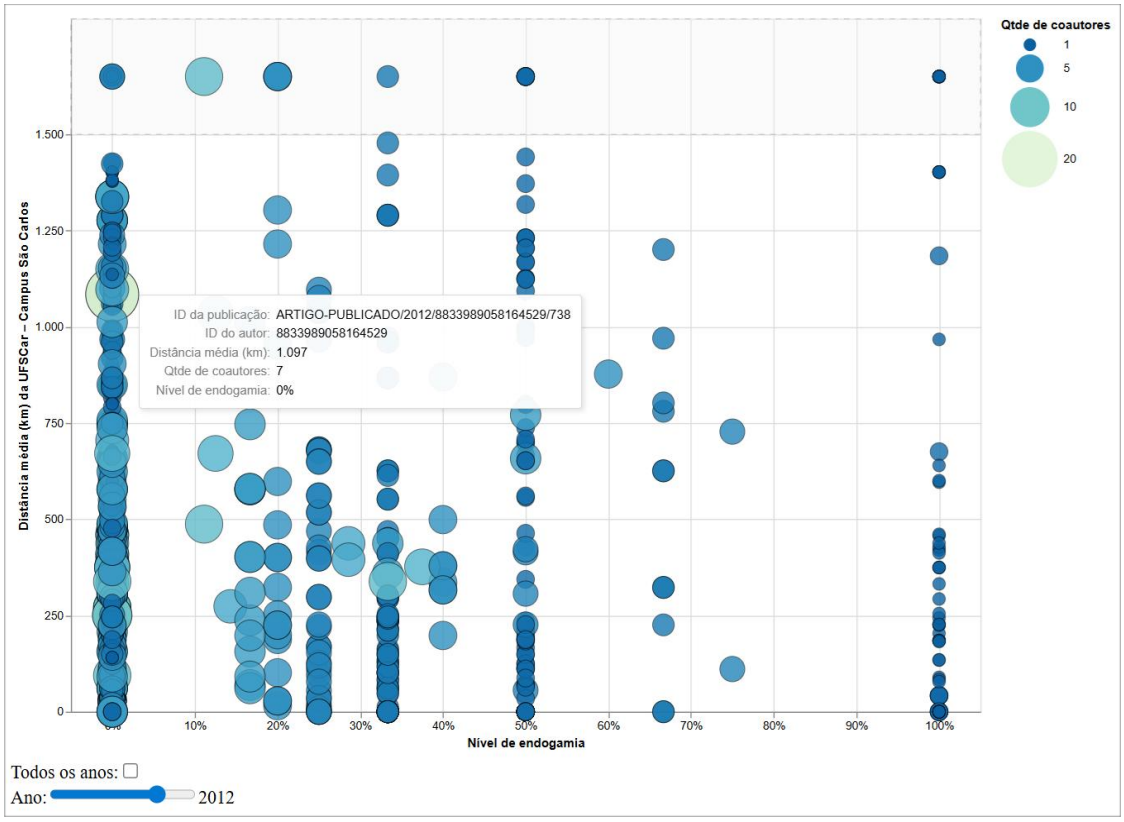
O gráfico de dispersão correlaciona o nível de endogamia às distâncias médias das coautorias, sob a ótica individual de cada pesquisador da UFSCar – Campus São Carlos. Desta forma, uma única publicação pode gerar múltiplos registros, sendo contabilizada distintamente para cada um dos seus coautores também vinculados à instituição. Por ser interativa, essa visualização é rica em

detalhes, o que potencializa a análise. Além de diferenciar o tamanho e a cor dos nós de acordo com a quantidade de coautores, ela retorna informações da publicação ao posicionar o *cursor* em cima do nó.

Na Figura 42 observa-se a presença expressiva de pesquisadores com baixos níveis de endogamia ao longo de diferentes faixas de distância média. A visualização evidencia a coexistência, em distâncias médias mais elevadas, de trajetórias com baixos e elevados níveis de endogamia, sendo estas últimas predominantemente associadas a nós de menor tamanho, isto é, trajetórias com reduzido número de coautores. Não se identifica, contudo, uma característica linear ou monotônica entre distância média e endogamia, indicando que a dimensão espacial, isoladamente, não é suficiente para explicar a incidência da endogamia nas trajetórias analisadas.

Para mostrar a interatividade proporcionada por essa visualização, na Figura 43 é apresentada a mesma visualização dinâmica, porém, com a utilização do filtro de ano.

Figura 43 – Visualização dinâmica do nível de endogamia e distância geográfica dos pesquisadores com filtro de ano.



Fonte: elaborada pelo autor.

Como pode ser visto, no filtro para seleção do ano a opção “Todos” está desabilitada e o ano de 2012 está selecionado. Isso significa que todas as publicações que aparecem nessa imagem são desse ano em específico.

As análises apresentadas neste capítulo permitiram caracterizar a endogamia nas coautorias científicas a partir de diferentes unidades de observação — coautorias, publicações, pesquisadores e instituições — evidenciando a heterogeneidade dos comportamentos de colaboração no conjunto de dados analisado. De modo geral, os resultados indicam que a endogamia apresenta ocorrência limitada no plano agregado, coexistindo com dinâmicas colaborativas predominantemente não endogâmicas.

A análise das coautorias e publicações mostrou o predomínio de relações não endogâmicas, ao passo que a endogamia, embora minoritária, manifesta-se de forma concentrada em casos específicos, variando conforme a unidade de análise adotada. Quando observada no nível das publicações e dos pesquisadores, a endogamia revela assimetria, com a maioria dos registros concentrando-se em níveis baixos, ao lado de casos pontuais com intensidades mais elevadas.

No plano institucional, os resultados indicam que instituições com maior volume de produção tendem a apresentar baixos níveis percentuais de endogamia, enquanto níveis mais elevados desse fenômeno aparecem com maior frequência entre instituições com menor produção. No entanto, a análise conjunta entre endogamia, volume de produção e distância geográfica não evidenciou associações lineares que permitam atribuir à proximidade espacial um papel determinante na intensificação da endogamia institucional.

As visualizações que relacionam colaboração institucional e distância geográfica indicam que a proximidade espacial está associada a maiores volumes absolutos de coautorias e publicações, sem que isso implique, necessariamente, maiores níveis de endogamia. Da mesma forma, a distribuição das publicações não endogâmicas ao longo do eixo espacial não apresenta associação direta com a distância geográfica, sugerindo que esse tipo de produção se distribui de forma heterogênea no espaço.

A análise no nível dos pesquisadores, incorporando simultaneamente o nível de endogamia, a distância média das coautorias e o volume relativo de colaboração,

evidenciou a presença expressiva de trajetórias caracterizadas por baixos níveis de endogamia em diferentes faixas de distância média. Observou-se, ainda, a coexistência de pesquisadores com baixos e elevados níveis de endogamia ao longo do espectro de distâncias analisado, sendo os níveis mais elevados frequentemente associados a trajetórias com menor número de coautores. Não se identificou, contudo, uma relação linear ou monotônica entre distância média e endogamia.

Em conjunto, os resultados indicam que a endogamia nas coautorias científicas não pode ser explicada de forma satisfatória por variáveis isoladas, como a distância geográfica ou o volume de produção. Ao contrário, observou-se que a endogamia emerge de dinâmicas relacionais e institucionais mais complexas. Esses achados fornecem subsídios empíricos para a discussão teórica apresentada ao longo da tese.

Por fim, destaca-se que as visualizações dinâmicas geradas nesta pesquisa, em formato HTML, encontram-se disponíveis no repositório de dados Zenodo[42], permitindo a manipulação interativa dos dados diretamente no navegador Web, indo ao encontro dos conceitos de InfoVis que foram apresentados.

[42] Visualizações dinâmicas para analisar a endogamia das coautorias científicas dos pesquisadores da UFSCar – Campus São Carlos credenciados em PPGs da universidade. Disponível em: <https://zenodo.org/records/18446461>. Acesso em: 01 fev. 2026.

5 CONSIDERAÇÕES

O processo de construção do conhecimento e da ciência está relacionado ao potencial de divulgação e publicação de 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, que são essenciais para a compreensão dos impactos das políticas e das dinâmicas sociais envolvendo ciência e tecnologia.

A endogamia acadêmica possui alguns potenciais problemas. A redução da diversidade de pensamento que ocorre no compartilhamento dos mesmos antecedentes e experiências acadêmicas pode levar à diminuição da qualidade da produção científica e diminuir o potencial de inovação. Um ambiente acadêmico fechado dificulta a entrada de novas ideias e perspectivas, o que pode induzir a um declínio na produtividade científica. Há também implicações de natureza ética, uma vez que, se a endogamia é alta, ainda que não intencionalmente, poderá estar ocorrendo favorecimentos de candidatos que foram formados na própria instituição.

Nesse contexto, a PL consolida-se como uma fonte de dados essencial para a elaboração de indicadores de ciência e tecnologia. Ao responder o problema central desta pesquisa — que analisou a endogamia acadêmica em coautorias de pesquisadores da UFSCar – Campus São Carlos credenciados em PPGs — confirmou-se que é possível compreender as características de endogamia e proximidade geográfica a partir de dados da referida plataforma, desde que submetidos a métodos estruturados de tratamento e visualização. A aplicação do procedimento computacional desenvolvido mostrou a viabilidade técnica de utilizar um *framework* de serialização de objetos a partir de documentos JSON, para converter registros brutos em informações geoespaciais e relacionais explícitas.

As análises realizadas permitiram confirmar a hipótese desta pesquisa, mostrando que métodos de tratamento e visualização de dados tornam explícitas as relações de colaboração que estão ocultas na PL. Sob essa nova perspectiva, relevou-se que a endogamia acadêmica não é dominante no conjunto das coautorias e publicações analisadas. No plano agregado, predominam relações colaborativas não endogâmicas, indicando um cenário caracterizado por circulação e diversificação das parcerias científicas. A endogamia, embora presente e agora

devidamente mapeada pela metodologia proposta, manifesta-se de forma minoritária e concentrada em casos específicos, assumindo diferentes intensidades conforme a unidade de análise considerada.

Sob a perspectiva da sociologia da ciência de Bourdieu, os resultados desta tese permitem interpretar as redes de coautoria na UFSCar como uma representação da estrutura do campo científico, onde o capital social — acumulado por meio de parcerias e colaborações — desempenha um papel central na posição ocupada pelos agentes. A identificação de que a endogamia é um fenômeno minoritário sugere que, embora a proximidade geográfica facilite a interação física, os pesquisadores buscam a diversificação de suas redes para ampliar seu capital simbólico e evitar o isolamento acadêmico. Por outro lado, a persistência de núcleos endogâmicos, mesmo que reduzidos, pode indicar estratégias de conservação de posições dentro do campo, onde a vantagem cumulativa mencionada por Merton se traduz em uma estrutura de posições consolidada pela proximidade e pelo compartilhamento de recursos locais.

A incorporação da dimensão espacial permitiu aprofundar a análise dessas dinâmicas, indicando que a proximidade geográfica está associada a maiores volumes absolutos de coautorias e publicações, mas não se apresenta como fator explicativo isolado da endogamia acadêmica. As visualizações não evidenciaram associações lineares ou monotônicas entre distância geográfica e níveis de endogamia, tanto no nível institucional quanto no nível dos pesquisadores. Observa-se a coexistência de diferentes níveis de endogamia ao longo de todo o espectro espacial analisado.

Em conjunto, os resultados indicam que a endogamia acadêmica nas coautorias científicas emerge de dinâmicas relacionais e institucionais complexas, não podendo ser explicada de forma satisfatória a partir de variáveis isoladas, sejam elas espaciais ou quantitativas. Esses achados reforçam a importância de abordagens integradas, que considerem simultaneamente múltiplas escalas de análise e explorem o potencial analítico das visualizações da informação para o estudo das redes científicas.

Como contribuição, esta tese apresenta um procedimento metodológico replicável para análise de coautorias científicas a partir da PL, bem como um

conjunto de visualizações que ampliam a capacidade interpretativa dos dados bibliométricos. A pesquisa contribui para o debate sobre coautoria científica, endogamia acadêmica e visualização de informação no contexto dos estudos métricos da ciência.

Reconhecem-se, por fim, limitações inerentes às escolhas metodológicas adotadas, que não esgotam a complexidade do fenômeno analisado. Essas limitações apontam para possibilidades de pesquisas futuras, incluindo análises comparativas entre instituições, áreas do conhecimento ou contextos nacionais distintos, bem como a incorporação de abordagens qualitativas que aprofundem a compreensão dos fatores institucionais e relacionais associados à endogamia acadêmica.

REFERÊNCIAS

ALTBACH, P. G.; YUDKEVICH, M.; RUMBLEY, L. E. Academic inbreeding: local challenge, global problem. **Asia Pacific Education Review**, p. 317-330, 2015. Disponível em: <https://link.springer.com/article/10.1007/s12564-015-9391-8>. Acesso em: 25 ago. 2023.

ALVES, M. C. **Visualização de informação para simplificar o entendimento de indicadores sobre avaliação da ciência e tecnologia**. 2015. 120 f. Dissertação (Mestrado em Ciência, Tecnologia e Sociedade) - Universidade Federal de São Carlos, São Carlos, 2015.

ALVES, M. C.; FARIA, L. I. L.; AMARAL, R. M. Visualização de informação para simplificar o entendimento de indicadores sobre avaliação da ciência e tecnologia. **Revista Digital Biblioteconomia e Ciência da Informação (RDBCI)**. Campinas, SP, v.15, n.2, p. 324-348, 2017. DOI: 10.20396/rdbci.v15i2.8646366.

ALZAMORA, P. L. *et al*. A COVID-19 no Twitter: correlacionando vocabulário com agravamento e atenuação da pandemia no Brasil. In: XI Brazilian Workshop on Social Network Analysis and Mining (BRASNAM), 2022, Niterói. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2022. p. 157-168. ISSN 2595-6094. DOI: 10.5753/brasnam.2022.223330.

AMARAL, R. M.; MATIAS, M. S. O.; SARVO, D. O. Interdisciplinaridade da Ciência da Informação brasileira: intensidades e relações. **Em Questão**, Porto Alegre, v. 30, 2024. Disponível em: <https://seer.ufrgs.br/index.php/EmQuestao/article/view/131695>. Acesso em: 16 nov. 2025.

AMORAS, F. C. EDITORIAL: Exogenia, endogenia e qualis das revistas. **Estação Científica (UNIFAP)**. ISSN 2179-1902. Macapá, v. 7, n. 3, p. 7-8, set./dez. 2017. DOI: 10.18468/estcien.2017v7n3.p07-08.

ANGELOZZI, S. M.; MARTÍN, S. G. **Metadatos para la descripción de recursos electrónicos en línea**: análisis y comparación. Buenos Aires: Alfagrama, 2010.

ARAKAKI, F. A.; SIMIONATO, A. C.; SANTOS, P. L. V. A. da C. Catalogação e tecnologia: interseções com a Web Semântica. **Informação@Profissões**, Londrina, v. 6, n. 2, p. 3-19, jul./dez. 2017.

ARAUJO, R. F. A altmetria na prática e o papel dos bibliotecários no seu uso e aplicação. **Em Questão**, Porto Alegre, v. 24, n. 1, p. 296-302, jan./abr. 2018.

AUTRAN, M. M. M. *et al*. Perfil de produção acadêmica dos programas brasileiros de pós-graduação em Ciência da Informação 2008-2012. **Perspectivas em Ciência da informação**, Belo Horizonte, v. 20, n. 4, p. 57-78, out./dez. 2015.

BAPTISTA, A. A.; MACHADO, A. B. Um gato preto num quarto escuro: falando sobre metadados. **Revista de Biblioteconomia de Brasília**, v. 25, n. 1, p. 77-90, 2001. Disponível em: <https://brapci.inf.br/index.php/res/v/77994>. Acesso em: 16 nov. 2025.

BAPTISTA, D. M. O impacto dos metadados na representação descritiva. **Revista ACB**: Biblioteconomia em Santa Catarina, v. 12, n. 2, p. 177-190, 2007. Disponível em: <https://brapci.inf.br/index.php/res/v/225943>. Acesso em: 22 ago. 2023.

BASSOLI, M. **Avaliação do Currículo Lattes como fonte de informação para construção de indicadores**: o caso da UFSCar. 2017. 84 p. Dissertação (Mestrado em Ciência, Tecnologia e Sociedade) - Universidade Federal de São Carlos, São Carlos, 2017.

BASTOS, M. T.; ZAGO, G.; RECUERO, R. A endogamia da Comunicação: redes de colaboração na CSAI. **Revista Famecos (Online)**, Porto Alegre, v. 23, n. 2, p. 1-27, maio/ago. 2016. DOI: 10.15448/1980-3729.2016.2.21459. Acesso em: 12 out. 2025.

BERELSON, B. **Graduate education in the United States**. New York: McGraw-Hill, 1960.

BOON, C. Y.; FOON, J. W. J. **Altmetrics is an indication of quality research or just hot topics**. Singapore: Nanyang Technological University, 2014. Disponível em: <https://www.researchgate.net/publication/264423415_Altmetrics_is_an_Indication_of_Quality_Research_or_Just_HOT_Topics>. Acesso em: 29 ago. 2025.

BORENSTEIN, D.; PERLIN, M. S.; IMASATO, T. The Academic Inbreeding Controversy: Analysis and Evidence from Brazil. **Journal of Informetrics 16**, 2022. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1751157722000396>. Acesso em: 26 ago. 2023.

BOSCHMA, R. A. Proximity and innovation: a critical assessment. **Regional Studies**, v. 39, n. 1, p. 61-74, fev. 2005. Disponível em: <https://doi.org/10.1080/0034340052000320887>. Acesso em: 03 jun. 2025.

BOURDIEU, P. O campo científico. In: ORTIZ, R. (Org). **Pierre Bourdieu**: sociologia. São Paulo: Ática, p. 122-155, 1983.

BOURDIEU, P. **Para uma sociologia da ciência**. Trad. Pedro Elói Duarte. Lisboa: Edições 70, 2004.

BOUTIN, E.; DUMAS, P.; ROSTAING, H.; QUONIAM, L. Les reseaux comme outils d'analyse en bibliométrie. Un cas d'application: les reseaux d'auteurs. **Cahier de la Documentation**, n. 1, p. 3-13, mars. 1996.

BRAGA, M. J. C.; GOMES, L. F. A. M.; RUEDIGER, M. A. Mundos pequenos, produção acadêmica e grafos de colaboração: um estudo de caso dos Enanpads. **Revista de Administração Pública**, Rio de Janeiro, 42(1):133-154, jan./fev. 2008. ISSN 0034-7612.

BRAGA, M. M. S; VENTURINI, A. E. J. F. Endogenia acadêmica em um programa de pós-graduação em direito. In: MEZZAROBA, O.; TAVARES-NETO, J. Q.; VASCONCELOS, S. A. (Coord.). **Direito, educação, ensino e metodologia jurídicos**. FUNJAB: Curitiba, 2013. p. 91-108.

BRINTON, W. C. **Graphic Presentation**. 1939. From the collection of the Prelinger Library, San Francisco, California, 2008.

BRITO, J. F.; MARTÍNEZ-ÁVILA, D. Metadados de preservação digital na era de Big Data. In: MARTÍNEZ-ÁVILA, D.; SOUZA, E. A.; GONZALEZ, M. E. Q. (Org.). **Informação, conhecimento, ação autônoma e big data**: continuidade ou revolução. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica; FiloCzar, 2019. ISBN: 978-85-7249-055-9.

BUSH, V. As we may think. **Atlantic Monthly**, v. 176, n. 1, p. 101-108, 1945. Disponível em: <http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm>. Acesso em: 22 ago. 2023.

CAMARGO JR, K. R.; COELI, C. M. Múltipla autoria: crescimento ou bolha inflacionária? **Rev. Saúde Pública**, p. 894-900, 2012.

CAMPEROS-REYES, J. T.; ROMANETTO, L. M.; SANT'ANA, R. C. G.; SANTOS, P. L. V. A. C. Intersecção temática de programas de pós-graduação brasileiros: considerações sobre Big Data. In: MARTÍNEZ-ÁVILA, D.; SOUZA, E. A.; GONZALEZ, M. E. Q. (Org.). **Informação, conhecimento, ação autônoma e big data**: continuidade ou revolução. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica; FiloCzar, 2019. ISBN: 978-85-7249-055-9.

CARD, S. K.; MACKINLAY, J. D.; SHNEIDERMAN, B. **Readings in information visualization**: using vision to think. São Francisco: Morgan Kaufmann, 1999. 686 p.

CATANI, A. M.; NOGUEIRA, M. A.; HEY, A. P.; MEDEIROS, C. (Orgs.). **Vocabulário Bourdieu**. Belo Horizonte: Autêntica, 2017.

CENTRO DE GESTÃO E ESTUDOS ESTRATÉGICOS (CGEE). **Brasil**: Mestres e doutores 2019. Brasília, DF: CGEE, 2021. Disponível em: <https://mestresdoutores2019.cgee.org.br>. Acesso em: 27 ago. 2023.

CIÊNCIA SEM FRONTEIRAS (CSF). **Programa Ciência Sem Fronteiras**. 2022. Disponível em: <https://www.gov.br/cnpq/pt-br/acesso-a-informacao/acoes-e-programas/programas/ciencia-sem-fronteiras/apresentacao-1/o-que-e>. Acesso em: 01 set. 2023.

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO (CNPq). **Plataforma Lattes**. Brasília, DF: CNPq, 2023. Disponível em: <http://lattes.cnpq.br>. Acesso em: 21 ago. 2023.

COSTAS, R. Discussões gerais sobre as características mais relevantes de infraestruturas de pesquisa para a cientometria. In: Mugnaini, R.; Fujino, A.; Kobashi, N. Y. (Org). **Bibliometria e cientometria no Brasil**: infraestrutura para avaliação da pesquisa científica na era do Big Data. São Paulo, ECA – USP, 2017. DOI: 10.11606/9788572051705.

COTTA, E. M. S.; DELBIANCO, N. R.; HILÁRIO, C. M. Infometria para sistemas de recuperação de informação. In: GRÁCIO, M. C. C. *et al*. (Org.). **Tópicos da bibliometria para bibliotecas universitárias**. Marília: Oficina Universitária, São Paulo: Cultura Acadêmica, p. 192-207. 2020.

CROSS, R.; PARKER, A.; SASSON, L. **Networks in the Knowledge Economy**. Oxford University Press. 2003.

DAMACENO, R. J. P; HADDAD, E. A.; MENA-CHALCO, J. P. Formação, endogenia e influência institucional na academia brasileira: uma análise da absorção de doutores nas instituições de ensino superior. 6° Encontro Brasileiro de Bibliometria e Cientometria (EBBC), **Anais** [...], Rio de Janeiro, 2018.

DAVENPORT, Thomas H. **Big data at work**: dispelling the myths, uncovering the opportunities. Harvard: Harvard Business School Publishing, 2014.

DAVYT, A.; VELHO, L. A avaliação da ciência e a revisão por pares: passado e presente. Como será o futuro? **História, ciência, saúde**. Manguinhos, Rio de Janeiro, v. 7, n. 1, p. 93-116, mar./jun. 2000.

DHAR, V. **Data Science and Prediction**. CACM 56, p. 12, 2013.

DIAS, T. M. R. *et al*. Integração de repositórios de dados abertos para certificação de produção científica. **BiblioCanto**, Natal, v. 9, n. 2, p. 12-16, 2023. DOI: 10.21680/2447-7842.2023v9n2ID33837.

DIAS, T. M. R.; MOITA, G. F.; DIAS, P. M. Um estudo sobre a rede de colaboração científica dos pesquisadores brasileiros com currículos cadastrados na Plataforma Lattes. **Em Questão**, Porto Alegre, v. 25, n. 1, p. 63-86, jan./abr. 2019. DOI: 10.19132/1808-5245251.63-86.

DORTA-GONZÁLEZ, P.; DORTA-GONZÁLEZ, M. I. Indicador bibliométrico basado en el índice h. **Revista española de Documentación Científica**. Madrid, v. 33, n. 2, p. 225-245, 3 maio 2010. Disponível em: <http://redc.revistas.csic.es/index.php/redc/article/view/553/627>. Acesso em: 22 ago. 2023.

DUMBILL, E. **Planning for Big Data**. O'Reilly Media. 2012.

ENGWALL, L.; BLOCKMANS, W.; WEAIRE, D. **Bibliometrics**: issues and context. Portland Press Limited. 2014. Disponível em: <https://portlandpress.com/DocumentLibrary/Umbrella/Wenner%20Gren/Vol%2087/WG_87_chapter%201.pdf>. Acesso em: 25 ago. 2023.

FARIA, L. I. L. *et al*. Análise da produção científica a partir de publicações em periódicos especializados. In: BRENTANI, R. R.; CRUZ, C. H. B.; SUZIGAN, W.; FURTADO, J. E. M. P.; GARCIA, R. C. (Org.). **Indicadores de Ciência, Tecnologia e Inovação em São Paulo - 2010**. 1 ed. São Paulo: FAPESP, 2011, v. 1, p. 1-71.

FERREIRA, P. G.; MCMANUS, C. M.; FARIA, L. I. L. Programas de cooperação acadêmica internacional e pesquisas colaborativas: resultados e tendências. **Inf. Inf.**, Londrina, v. 27, n. 3, p. 535-556, jul./set. 2022. DOI: 10.5433/1981-8920.2022v27n3p535.

FEW, S. **Information Dashboard Design**: The Effective Visual Communication of Data. O'Reilly, 2006.

FOREMAN, J. W. **Data Smart**: Using Data Science to Transform Information into Insight. John Wiley & Sons, Hoboken, NJ. 2013.

FRANCO, N. M. G.; FARIA, L. I. L. Colaboração científica intraorganizacional: análise de redes por coocorrência de palavras-chave. **Em Questão**, Porto Alegre, v. 25, n. 1, p. 87-110, jan./abr. 2019. DOI: 10.19132/1808-5245251.87-110.

FREIRE, G. H. A. Ciência da Informação: temática, histórias e fundamentos. **Perspectivas em Ciência da Informação**, v. 11, n.1, p. 6-19, 2006.

FREITAS, C. M. D. S. *et al*. Introdução à visualização de informações. **RITA**, Porto Alegre, v. 8, n. 2, p. 143-158, 2001.

FREITAS, E. R. S.; BENCHIMOL, A. C.; RODRIGUES, F. A. Análise de publicações do Museum Week 2022-2023 no serviço de rede social online Twitter. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO –

ENANCIB, 23., 2023, Aracaju. **Anais** [...]. Aracaju: ANCIB, 2023. p. 1-18. Disponível em: <https://www.researchgate.net/publication/379515338>. Acesso em: 22 ago. 2025.

FELIPE, L. L. *et al.* Liderança e colaboração internacional na pesquisa sobre Covid-19. 8° Encontro Brasileiro de Bibliometria e Cientometria (EBBC), **Anais** […], Maceió, 2022. DOI: https://doi.org/10.22477.110.

FUNDAÇÃO OSWALDO CRUZ (FIOCRUZ). **Observatório da Fiocruz**. Rio de Janeiro: FIOCRUZ, 2023. Disponível em: <https://observatorio.fiocruz.br>. Acesso em: 24 ago. 2023.

FURTADO, C. A.; DAVIS JR, C. A.; GONÇALVES, M. A.; ALMEIDA, J. M. **A spatio temporal analysis of Brazilian science from the perspective of researchers career trajectories**. PloS One, 2015. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0141528>. Acesso em: 25 ago. 2023.

GAZDA, E.; QUANDT, C. O. Colaboración interinstitucional en la investigación en brasil: Tendencias de los artículos en el área de gestión de innovación. **RAE-eletrônica**, v. 9, n. 2, 2010.

GIL, A. C. **Como elaborar projetos de pesquisa**. 6. ed. São Paulo: Atlas, 2017.

GLÄNZEL, W.; THIJS, B.; DEBACKERE, K. Citation Classes: A Distribution-based Approach for Evaluative Purposes. In: GLÄNZEL, W.; MOED, H. F.; SCHMOCH, U.; THELWALL, M. (Org.). **Springer Handbook of Science and Technology Indicators**. Springer Nature Switzerland AG, 2019.

GODECHOT, O.; LOUVET, A. Inbreeding in Universities: In Favour of Administrative Regulation. **La Vie des Idées**, Paris, 13 mai. 2008. Tradução do Francês para o Inglês por Susannah Dale.

GONTIJO, M. C. A.; ARAÚJO, R. F. Impacto acadêmico e atenção on-line de pesquisas sobre inteligência artificial na área da saúde: análise de dados bibliométricos e altmétricos. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 26, p. 1-21, 2021. DOI: https://doi.org/10.5007/1518-2924.2021.e76249.

GOOGLE. **Google Maps Platform**: Pricing and Plans. Google Cloud, 2025. Disponível em: <https://cloud.google.com/maps-platform/pricing>. Acesso em: 08 nov. 2025.

GOUVEIA, F. C.; ARAÚJO, R. F. Webometria: origens e usos contemporâneos. In: GRÁCIO, M. C. C. *et al*. (Org.). **Tópicos da bibliometria para bibliotecas universitárias**. Marília: Oficina Universitária, São Paulo: Cultura Acadêmica, p. 208-228. 2020.

GRACIOSO, L. de S. Consumo e uso da informação na Web: pragmática informacional na modernidade líquida. **Informação**: agentes e intermediação, p. 355-390, Brasília/DF: IBICT, 2017.

GREGOLIN, J. A. R. *et al*. Análise da produção científica a partir de indicadores bibliométricos. In: LANDI, F. R.; GUSMÃO, R. (Org.). **Indicadores de ciência, Tecnologia e Inovação em São Paulo-2004**. Fapesp, São Paulo, v. 1, p. 1-44, 2005.

GROCHOCKI, L. F. M.; CABELLO, A. F. Academic endogamy or immobility? The impact on scholarly productivity in a developing country. **International Journal of Educational Development**, v. 94, 2022. DOI: 10.1016/j.ijedudev.2022.102652.

GROENNER, L. C. *et al*. Um Estudo Bibliométrico sobre a pesquisa em Inteligência Artificial no Brasil. **Brazilian Journal of Information Science**: Research trends, vol. 16, 2022. Disponível em: <https://revistas.marilia.unesp.br/index.php/bjis/article/view/12855/8653>. Acesso em: 09 nov. 2025.

GUERRA, A. L. R.; AUGUSTO, E. A.; LEÃO, U. D. F. Mobilidade internacional na pós-graduação brasileira: impactos acadêmicos e profissionais do doutorado sanduíche. **Revista Eletrônica Multidisciplinar de Investigação Científica**, v. 4, n. 22, p. 190-201, 2025. DOI: 10.56166/remici.v4n22171725. Acesso em: 12 out. 2025.

GUIMARÃES, D. B.; ROCHA, E. S. S.; MUGNAINI, R. Estudo cientométrico da atividade acadêmica sobre as temáticas de humanidades digitais e big data nas universidades estaduais paulistas. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 28, p. 1-22, 2023. Disponível em: <https://doi.org/10.5007/1518-2924.2023.e90566>. Acesso em: 26 maio 2025.

GUSMÃO, A. C. S.; MENA-CHALCO, J. P. Analisando a hipercoautoria científica: um estudo de caso com a CMS Collaboration (CERN). In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB, 24., 2024, Vitória. **Anais** [...]. Vitória: ANCIB, 2024. Disponível em: < https://www.researchgate.net/publication/386246085_ANALISANDO_A_HIPERCOAUTORIA_CIENTIFICA_UM_ESTUDO_DE_CASO_COM_A_CMS_COLLABORATION_CERN>. Acesso em: 22 ago. 2025.

GUSMÃO, A. C. S.; SANTOS, S. M.; MENA-CHALCO, J. P. Análise da longevidade e do tamanho das coautorias acadêmicas: os caminhares na ciência brasileira. **Em Questão**, Porto Alegre, v. 28, n. 2, abr./jun. 2022. DOI: https://doi.org/10.19132/1808-5245282.116156. Acesso em: 26 ago. 2025.

HEAD, K.; LI, Y. A.; MINONDO. A. Geography, ties, and knowledge flows: Evidence from citations in mathematics. **Review of Economics and Statistics**, 2018. DOI: 10.1162/rest_a_00771.

HERZOG, C.; HOOK, D.; KONKIEL, S. Dimensions: Bringing down barriers between scientometricians and data. **Quantitative Science Studies**, v. 1, n. 1, p. 387-395, 2020. DOI: https://doi.org/10.1162/qss_a_00020.

HILÁRIO, C. M.; FREITAS, J. L. Indicadores de colaboração científica: aspectos éticos, práticos e formas de mensuração. In: GRÁCIO, M. C. C. *et al*. (Org.). **Tópicos da bibliometria para bibliotecas universitárias**. Marília: Oficina Universitária, São Paulo: Cultura Acadêmica, p. 72-93. 2020.

HILLMANN, D. **Using Dublin Core**. 2005. Disponível em: <http://dublincore.org/documents/usageguide>. Acesso em: 22 ago. 2023.

HOEREN, T.; KOLANY-RAISER, B. **Big Data in Context**: Legal, Social and Technological Insights (SpringerBriefs in Law). Springer International Publishing. 2017.

HORTA, H. Deepening our understanding of academic inbreeding effects on research information exchange and scientific output: new insights for academic based research. **Higher Education**, v. 65, n. 4, p. 487-510, abr. 2013.

HORTA, H.; MEOLI, M.; SANTOS, J. M. Academic inbreeding and choice of strategic research approaches. **Higher Education Quarterly**, 2021. DOI: 10.1111/hequ.12328.

HORTA, H.; YUDKEVICH, M. The role of academic inbreeding in developing higher education systems: challenges and possible solutions. **Technological Forecasting and Social Change**. 113: 363-372. 2016.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA (IBICT). **BrCris**. Brasília, DF: IBICT, 2023. Disponível em: <https://brcris.ibict.br>. Acesso em: 24 ago. 2023.

JOHN, M.; FRITSCHE, F. **Bibliometrics for Technology Forecasting and Assessment**. Fraunhofer Institute for Technological Trend Analysis, Euskirchen, 2013.

JOHNSON, M. **Geodesy and Navigation**. New York: Academic Press, 2015.

JUSTINO, T. S. **Análise da colaboração científica dos programas de Pós-graduação em Ciência da Informação brasileiros**. 2019. 103 p. Dissertação (Mestrado em Ciência da Informação) - Universidade Federal de São Carlos, São Carlos, 2019.

KEIM, D. A. Information Visualization and Visual Data Mining. **IEEE Transactions on Visualization and Computer Graphics**, v. 7, n. 1, p. 100-107, 2002.

KESSLER, M. M. **Bibliographic coupling between scientific papers**. American Documentation, v. 14, p. 10-25, 1963.

KRACKHARDT, D. Assessing the Political Landscape: Structure, Cognition, and Power in Organizations. **Administrative Science Quarterly**, [s. l.], v. 35, n. 2, p. 342-369, 1990.

KRACKHARDT, D. Graph Theoretical Dimensions of Informal Organizations. In: CARLEY, K.; PRIETULA, M. (eds.). **Computational Organizational Theory**. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., p. 89-111, 1994.

LIMA, G. A. B. O. Organização e representação do conhecimento e da informação na web: teorias e técnicas. **Perspectivas em Ciência da Informação**, v. 25, p. 57-97, 2020. Disponível em: <http://hdl.handle.net/20.500.11959/brapci/135734>. Acesso em: 01 set. 2023.

LIMA, R. A. **Análise bibliométrica da atividade científica em bioprospecção (1986 - 2006)**. 2007. Dissertação (Mestrado) - Instituto de Geociências, Universidade Estadual de Campinas, Campinas, 2007.

LIMA, R. A.; VELHO, L. M. L. S.; FARIA, L. I. L. de. Indicadores bibliométricos de cooperação científica internacional em bioprospecção. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 12, n. 1, p. 50-64, jan./abr. 2007.

LOPES, A. C.; COSTA, H. H. C. A produção bibliográfica em coautoria na área de educação. **Revista Brasileira de Educação**, v. 17, n. 51, 2012.

LYU, X.; COSTAS, R. How do academic topics shift across altmetric sources? A case study of the research area of Big Data. **Scientometrics**, v. 123, p. 909-943, 2020. DOI: https://doi.org/10.1007/s11192-020-03415-7.

MACIEL, R. S. **A Plataforma Lattes como recurso estratégico para a gestão dos Programas de Pós-Graduação**: uma análise baseada na produção de artigos científicos. 2018. 183 p. Dissertação (Mestrado em Ciência da Informação) - Universidade Federal de São Carlos, São Carlos, 2018.

MACIEL, R. S.; FARIA, L. I. L.; MILANEZ, D. H.; LANÇA, T. A. Efeito Qualis e a produção científica dos programas de Pós-Graduação da Universidade Federal de São Carlos. **Em Questão**, Porto Alegre, v. 24, p. 88-110, Edição Especial do 6° Encontro Brasileiro de Bibliometria e Cientometria (EBBC), 2018. DOI: 10.19132/1808-5245240.88-110.

MARQUES, L. F. S.; SAYÃO, L. F. Conectando a eScience à Ciência da Informação: o big metadado científico e suas funcionalidades. **RDBCI**: Revista Digital de Biblioteconomia e Ciência da Informação, Campinas, v. 21, e023017, 2023. DOI: 10.20396/rdbci.v21i00.8673740.

MARQUES, R. S.; RODRIGUES, W. B.; DAMACENO, R. J. P.; MENA-CHALCO, J. P. A descendência acadêmica do Prof. César Lattes: uma caracterização em homenagem ao centenário de seu nascimento. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB. **Anais** [...]. Vitória: ENANCIB, 2024. p. 1-10.

MARTINHO, C. **Redes**: uma introdução às dinâmicas da conectividade e da auto-organização. 1. ed. Brasília: WWF-Brasil, 2003.

MATIAS, M. S. O. **Base referencial para o povoamento de repositórios institucionais**: coleta automatizada de metadados da Plataforma Lattes. 2015. 86 p. Dissertação (Mestrado em Gestão de Organizações e Sistemas Públicos) - Universidade Federal de São Carlos, São Carlos, 2015.

MATIAS, M. S. O.; AMARAL, R. M.; MATIAS, P. **Proxy customizado para acesso ao web service da Plataforma Lattes**: apresentação. São Carlos: Universidade Federal de São Carlos, 2017. Disponível em: <https://www.researchgate.net/publication/317348485_Proxy_customizado_para_acesso_ao_web_service_da_Plataforma_Lattes_Presentation>. Acesso em: 07 jun. 2025.

MENA-CHALCO, J. P.; CESAR-JR, 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.

MERTON, Robert K. **Ensaios de sociologia da ciência**. São Paulo: Editora 34, 2013.

MOMENI, F. *et al*. The many facets of academic mobility and its impact on scholars' career. **Journal of Informetrics**, v. 16, 2022. DOI: https://doi.org/10.1016/j.joi.2022.101280.

NARIN, F.; OLIVASTRO, D.; STEVENS, K. S. Bibliometric theory, practice and problem. **Evaluation Review**, Thousand Oaks, California, v. 18, n. 1, p. 65-76, 1994.

NASCIMENTO, A. G. **Altmetria para bibliotecários**. Rio de Janeiro: Ciência Moderna, 2016.

NOMINATIM. **Nominatim API Documentation**. OpenStreetMap Foundation, 2025. Disponível em: <https://nominatim.org>. Acesso em: 08 nov. 2025.

NORMAN, D. A. **O design do dia a dia**. Rio de Janeiro, RJ: Rocco, 2006.

OKUBO, Y. Bibliometric indicators and analysis of research systems: methods and examples. **OECD Science, Technology and Industry Working Papers**, Paris, v. 97, n. 41, 1997/1, OECD, 1997.

OLIVA, G. *et al*. Ciência básica e aplicada para o futuro das cidades. In: ANDRICOPULO, A. D.; BONAGAMBA, T. J. (org.). **Ciência, tecnologia, inovação e o futuro de São Carlos**. São Carlos: Fundação Pró-Memória de São Carlos, 2023. p. 12-27. ISBN 978-65-89494-07-2.

OLIVEIRA, E. F. T.; CASTANHA, R.; GABRIEL JUNIOR, R. F.; BUFREM, L. S. Contexto da produção científica de Big Data: análise cientométrica. In: MARTÍNEZ-ÁVILA, D.; SOUZA, E. A.; GONZALEZ, M. E. Q. (Org.). **Informação, conhecimento, ação autônoma e big data**: continuidade ou revolução. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica; FiloCzar, 2019. ISBN: 978-85-7249-055-9.

OLIVEIRA, M. A. Acesso aberto e colaboração científica na pandemia de COVID-19. **Revista Brasileira de Informação Científica**, v. 10, n. 2, p. 45-60, 2020.

OLIVEIRA, T. M.; AMARAL, L. Políticas Públicas em Ciência e Tecnologia no Brasil: desafios e propostas para utilização de indicadores na avaliação. In: Mugnaini, R.; Fujino, A.; Kobashi, N. Y. (Org). **Bibliometria e Cientometria no Brasil**: infraestrutura para avaliação da pesquisa científica na Era do Big Data. São Paulo, ECA – USP, 2017. DOI: 10.11606/9788572051705.

OTTONICAR, S. L. C.; ATAYDE, G. R.; SANTA-EULALIA, L. A. O Big Data no desenvolvimento da indústria 4.0: novas perspectivas para o empreendedorismo acadêmico. In: MARTÍNEZ-ÁVILA, D.; SOUZA, E. A.; GONZALEZ, M. E. Q. (Org.). **Informação, conhecimento, ação autônoma e big data**: continuidade ou revolução. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica; FiloCzar, 2019. ISBN: 978-85-7249-055-9.

PATIL, D. **Building Data Science Teams**. O'Reilly Media. 2011.

PELEGRINI, T.; FRANÇA, M. T. A. Endogenia acadêmica: insights sobre a pesquisa brasileira. **Estud. Econ.**, São Paulo, vol. 50, n. 4, p. 573-610, out.-dez. 2020.

PENFIELD, T.; BAKER, M. J.; SCOBLE, R.; WYKES, M. C. Assessment, evaluations, and definitions of research impact: A review. **Research Evaluation**, v. 23, p. 21-32, 2014. Disponível em: <https://academic.oup.com/rev/article/23/1/21/2889056>. Acesso em: 25 ago. 2023.

PEREIRA, A. **Geometria da Terra**: Esferoide e suas Implicações. Rio de Janeiro: Ciência Moderna, 2012.

PERLIN, M. S. *et al*. The Brazilian scientific output published in journals: A study based on a large CV database. **Journal of Informetrics**, v. 11, p. 18-31, 2017. DOI: 10.1016/j.joi.2016.10.008.

PRADO, M. A. R.; CASTANHA, R. C. G. Indicadores: conceitos fundamentais e importância em CT&I. In: GRÁCIO, M. C. C. *et al*. (Org.). **Tópicos da bibliometria para bibliotecas universitárias**. Marília: Oficina Universitária, São Paulo: Cultura Acadêmica, p. 50-71. 2020.

PROVOST, F; FAWCETT, T. Data science and its relationship to Big Data and data-driven decision making. **Big Data**. Vol. 1, No. 1, p. 51-59, 2013.

PUERTA-DÍAZ, M.; MARTÍ-LAHERA, Y.; MARTÍNEZ-ÁVILLA, D. Altmetria: métricas alternativas para bibliotecários. In: GRÁCIO, M. C. C. *et al*. (Org.). **Tópicos da bibliometria para bibliotecas universitárias**. Marília: Oficina Universitária, São Paulo: Cultura Acadêmica, p. 230-262. 2020.

REIS, J. E. **Incipiência da disponibilidade de indicadores bibliométricos e altmétricos nos repositórios institucionais brasileiros**. 2016. 120 p. Dissertação (Mestrado em Ciência, Tecnologia e Sociedade) - Universidade Federal de São Carlos, São Carlos, 2016.

REIS, J. E. **Impacto da formação docente internacional na produção científica**: o caso da UFSCar. 2021. 153 p. Tese (Doutorado em Ciência, Tecnologia e Sociedade) - Universidade Federal de São Carlos, São Carlos, 2021.

REIS, J. E. *et al*. Impacto da formação docente no exterior na coautoria internacional: estudo da produção científica da Universidade Federal de São Carlos indexada na Web of Science. **Transinformação**, Campinas, v. 33, 2021.

RIBEIRO, A. E. Visualização de informação e alfabetismo gráfico: questões para a pesquisa. **Inf. & Soc.**: Est., João Pessoa, v.22, n.1, p. 39-50, jan./abr. 2012.

RIGHETTI, S. **Qual é a melhor? Origem, indicadores, limitações e impactos dos rankings universitários**. 2016. Tese (Doutorado em Política Científica e Tecnológica) – Instituto de Geociências, Universidade Estadual de Campinas, Campinas, 2016.

RIPOLI, S. C. C. *et al*. Inovação com conexão: uma jornada interdisciplinar na ideação da Plataforma Tecnológica InfoHub. In: CAMPEROS-REYES, Jacquelin Teresa *et al*. (org.). **I Seminário Internacional Informação, Conhecimento e Digitalidade para o Desenvolvimento Sustentável da Amazônia**. Belém: Universidade Federal do Pará, 2024. p. 32-41. Disponível em: <https://www.researchgate.net/publication/388954736>. Acesso em: 8 jun. 2025.

ROCCA, F. X. In Spain, Inbreeding Threatens Academe. **Chronicle of Higher Education**, 53, no. 22: 01-317. 2007.

RODRIGUES, F. A. **Estruturas de dados em serviços de redes sociais online**: uma abordagem metodológica de análise. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica, 2024. 312 p.

RODRIGUES, F. A., RODRIGUES, L. B. Mapeamento e categorização do termo rede social em comunicações científicas da ciência da informação no Brasil. **RICI**: R. Ibero-amer. Ci. Inf., Brasília, v. 16, n. 2, p. 329-345, maio./ago. 2023. ISSN 1983-5213. DOI: 10.26512/rici.v16.n2.2023.47935.

RODRIGUES, F. A.; SANT'ANA, R. C. G.. Privacy and Online Social Network: a model for analysis of collecting personal data. **Brazilian Journal of Information**

Science: research trends, vol. 17, publicação contínua, 2023, e023005. DOI: 10.36311/1981-1640.2023.v17.e023005.

RODRIGUES, W. B.; MENA-CHALCO, J. P. Redes de colaboração em bancas de defesa de doutorado: caracterização das relações além da tese. In: 9° Encontro Brasileiro de Bibliometria e Cientometria (EBBC), **Anais** [...], Brasília, 2024. DOI: 10.22477/ix.ebbc.312.

RODRIGUES, W. B.; TOMASSINI, C.; MENA-CHALCO, J. P. Análise da estrutura de redes de sugestão de pareceristas científicos. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB, 23., 2023, Aracaju. **Anais** [...]. Aracaju: ANCIB, 2023. Disponível em: <https://www.researchgate.net/publication/375231555>. Acesso em: 24 ago. 2025.

ROSSI, L.; DAMACENO, R. J. P.; MENA-CHALCO, J. P. Genealogia acadêmica: Um novo olhar sobre impacto acadêmico de pesquisadores. **Parcerias Estratégicas**. Brasília, DF, v. 23, n. 47, p. 197-212, jul-dez. 2018. Disponível em: <https://www.researchgate.net/publication/326960663_Genealogia_academica_Um_novo_olhar_sobre_impacto_academico_de_pesquisadores>. Acesso em: 30 ago. 2023.

ROSSI, L.; MENA-CHALCO, J. P. Caracterização de árvores de genealogia acadêmica por meio de métricas em grafos. In: **Anais do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)**, 2014. Disponível em: < https://sol.sbc.org.br/index.php/brasnam/article/view/6800>. Acesso em: 12 out. 2025.

SAMPAIO, H.; SANCHEZ, I. Formação acadêmica e atuação profissional de docentes em educação: USP e Unicamp. **Cadernos de Pesquisa**, São Paulo, v. 47, n. 166, p. 1268-1291, out./dez. 2017.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Inf. Inf.**, Londrina, v. 21, n. 2, p. 116-142, maio/ago. 2016. DOI: 10.5433/1981-8920.2016v21n2p116. Disponível em: <https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/27940/20124>. Acesso em: 29 ago. 2023.

SANT'ANA, R. C. G. Transdução informacional: impactos do controle sobre os dados. In: MARTÍNEZ-ÁVILA, D.; SOUZA, E. A.; GONZALEZ, M. E. Q. (Org.). **Informação, conhecimento, ação autônoma e big data**: continuidade ou revolução. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica; FiloCzar, 2019. ISBN: 978-85-7249-055-9.

SANTINI, P. H. **Painel de indicadores de desempenho para as IFES – Instituições Federais de Ensino Superior**: um modelo multidimensional. 2017. 159 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Pernambuco, Recife, 2017.

SANTOS, S. M. **O desempenho das universidades brasileiras nos rankings internacionais**: áreas de destaque da produção científica brasileira. 2015. 344 p. Tese (Doutorado em Ciência da Informação) - Universidade de São Paulo, São Paulo, ECA – USP, 2015.

SARVO, D. O. **Inteligência Acadêmica**: presença das universidades públicas estaduais paulistas na produção científica da UFSCar, UNIFESP e UFABC [Dataset].

Zenodo. 2023. DOI: 10.5281/zenodo.8299434. Disponível em: <https://zenodo.org/records/8299434>. Acesso em: 30 ago. 2023.

SARVO, D. O.; REIS, J. E.; AMARAL, R. M. Inteligência acadêmica a partir da Plataforma Lattes: presença das universidades públicas estaduais paulistas na produção científica da UFABC, UFSCar e UNIFESP. 8° Encontro Brasileiro de Bibliometria e Cientometria (EBBC), **Anais** […], Maceió, 2022.

SEDRAKYAN, G.; MANNENS, E.; VERBERT, K. Guiding the choice of learning dashboard visualizations: Linking dashboard design and data visualization concepts. **Journal of Computer Languages**, v. 50, p. 19-38, 2019. ISSN 2590-1184.

SHIBAYAMA, S. Development of originality under inbreeding: A case of life science labs in Japan. **Higher Educ Q**., 2022, 76:63-75. DOI: 10.1111/hequ.12315.

SHNEIDERMAN, B. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In: **Proceedings of IEEE Symposium on Visual Languages**, Boulder, CO, p. 336-343, 1996.

SIDONE, O. J. G.; HADDAD, E. A.; MENA-CHALCO, J. P. A ciência nas regiões brasileiras: evolução da produção e das redes de colaboração científica. **TransInformação**, Campinas, v. 28, n. 1, p. 15-31, 2016.

SILVA, L. L.; GOMES, A. G.; FONTELES, D. M.; RODRIGUES, F. A. Indexação Social como estratégia para visibilidade de conteúdos científicos em Serviço de rede social online Tiktok. **Organização e Representação do Conhecimento em diferentes contextos**: desafios e perspectivas na era da datificação, 2023.

SILVA, N. G.; GOMES, A. G.; RODRIGUES, F. A. Análise de User Interface em Web Sites e Perfis em Serviços de Rede Sociais Online de Arquivos Estaduais Brasileiros. **Ágora**: Arquivologia em debate, ISSN 2763-9045, Florianópolis, v. 33, n. 67, p. 1-26, jul./dez. 2023.

SISMONDO, S. **An Introduction to Science and Technology Studies**. Wiley-Blackwell, 2 ed. Hong Kong, 2010.

SIVAK, E.; YUDKEVICH, M. **University Inbreeding**: An Impact on Values, Strategies and Individual Productivity of Faculty Members. National Research University – Higher School of Economics, 2012. Disponível em: <https://ssrn.com/abstract=1996417>. Acesso em: 19 out. 2025.

SMALL, H. **Citation analysis**. In: MOED, H. F.; GLÄNZEL, W.;

SMITH, J. **Earth's Radius and Geophysical Measurements**. London: Geophysical Publishing, 2010.

SMITH, R. A. Coautoria e ética na produção científica. **Journal of Scientific Publishing**, v. 8, n. 1, p. 50-65, 2019.

SOARES, G. A. D.; SOUZA, C. P. R.; MOURA, T. W. Colaboração na produção científica na Ciência Política e na Sociologia brasileiras. **Revista Sociedade e Estado**, v. 25, n. 3, 2010.

SOARES, V. R. M. *et al*. Visualização da colaboração científica entre pesquisadores a partir de metadados da Plataforma Lattes. In: 7° Encontro Brasileiro de Bibliometria e Cientometria (EBBC), **Anais** [...], Salvador, 2020, p. 563–572. ISSN 2675-5939. Disponível em: <http://www.ebbc.ici.ufba.br>. Acesso em: 31 dez. 2025.

SPINAK, E. Indicadores cienciométricos. **Ciência da Informação**, Brasília, v. 27, n. 2, p. 141-148, 1998. Disponível em: <http://revista.ibict.br/ciinf/article/view/795/826>. Acesso em: 22 ago. 2023.

TARGINO, M. G. Comunicação científica: uma revisão de seus elementos básicos. **Informação & Sociedade**, v. 10, n. 2, p. 1-27, 2000.

TARGINO, M. G. Orientador ou tutor é autor? **Inf. Inf.**, Londrina, v. 15, n. esp, p. 145-156, 2010.

TAVARES, O. *et al*. Inbreeding and research collaborations in Portuguese higher education. **Higher Education Quarterly**, v. 76, n. 1, p. 102-115, 2022. DOI: 10.1111/hequ.12301.

TAXWEILER, R.; SELL, D.; PACHECO, R. A Framework for Analytical Demands of Alumni Capital. **Proceedings of the 24th European Conference on Knowledge Management**, ECKM 2023, p. 1329-1337, 2023.

TRIBUNAL DE CONTAS DA UNIÃO (TCU). **Portal TCU**. Brasília, DF: TCU, 2023. Disponível em: <https://portal.tcu.gov.br>. Acesso em: 20 ago. 2023.

VAN RAAN, A. Measuring Science: Basic Principles and Application of Advanced Bibliometrics. In: GLÄNZEL, W.; MOED, H. F.; SCHMOCH, U.; THELWALL, M. (Org.). **Springer Handbook of Science and Technology Indicators**. Springer Nature Switzerland AG, 2019.

VANTI, N. A. P. Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 152-162, maio/ago. 2002.

VANZ, S. A. S. **As redes de colaboração científica no Brasil**. 2009. Tese (Doutorado em Comunicação e Informação) - Faculdade de Biblioteconomia e Comunicação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

VIANA, L. C. S.; MENA-CHALCO, J. P.; DAMACENO, R. J. P.; NABUCO, O. **Boletim Pesquisadores, Produção Bibliográfica e Orientação Acadêmica**: edição especial – redes de cooperação em pesquisa. 1. ed. Rio de Janeiro: Ed. dos Autores, 2023. ISBN 978-65-00-83969-2.

VINCENTY, T. Direct and Inverse Solutions of Geodesics on an Ellipsoid. **Journal of Geodesy**, v. 49, n. 4, p. 245-262, 1975.

WYATT, S.; MILOJEVIĆ, S.; PARK, H., W.; LEYDESDORFF, L. Intellectual and Practical Contributions of Scientometrics to STS. In: FELT, U.; FOUCHÉ, R.; MILLER, C. A.; SMITH-DOERR, L. (Org.). **The Handbook of Science and Technology Studies**. 4 ed. London: The MIT Press Cambridge, 2017.

WOLFRAM, D. A pesquisa bibliométrica na era do big data: Desafios e oportunidades. In: Mugnaini, R.; Fujino, A.; Kobashi, N. Y. (Org). **Bibliometria e Cientometria no Brasil**: infraestrutura para avaliação da pesquisa científica na Era do Big Data. São Paulo, ECA – USP, 2017. DOI: 10.11606/9788572051705.

YAMAGUCHI, J. K. **Diretrizes para a escolha de técnicas de visualização aplicadas no processo de extração do conhecimento**. 2010. 182 p. Dissertação

(Mestrado em Ciência da Computação) - Universidade Estadual de Maringá, Maringá, 2010.

YAN, E.; SUGIMOTO, C. R. Institutional interactions: Exploring social, cognitive, and geographic relationships between institutions as demonstrated through citation networks. **Journal of the American Society for Information Science and Technology**, 2011, 62(8), p. 1498-1514.

ZANIN, A. **Definição de painel de indicadores de desempenho para instituições comunitárias de ensino superior**. 2014. Tese (Doutorado em Engenharia de Produção) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2014.

ZÂNIRO, D. L.; QUONIAM, L. Analysis of patent production in Brazil: A perspective from the Lattes platform. **Advanced Notes in Information Science**, [S. l.], v. 8, p. 166–190, 2025. DOI: 10.47909/978-9916-9331-4-5.114. Disponível em: <https://anis.pro-metrics.org/index.php/a/article/view/114>. Acesso em: 31 dez. 2025.

ZHU, Y.; XIONG, Y. Towards Data Science. **Data Science Journal**, p. 1-7, 2015. Disponível em: <http://dx.doi.org/10.5334/dsj-2015-008>. Acesso em: 24 ago. 2023.

ZIMMER, M.; PROFERES, N. J. A topology of Twitter research: disciplines, methods, and ethics. **Aslib Journal of Information Management**, 66(3):250-261, 2014. DOI: 10.1108/AJIM-09-2013-0083.

ZITT, M.; BASSECOULARD, E. Internationalisation in Science in the Prism of Bibliometric Indicators. 2004. In: MOED, H.F., GLÄNZEL, W., SCHMOCH, U. (ed.). **Handbook of Quantitative Science and Technology Research**. Springer, Dordrecht. DOI: https://doi.org/10.1007/1-4020-2755-9_19.

APÊNDICE A – Conversor de arquivo XML para JSON.

Este *script* tem como objetivo converter arquivos XML provenientes da PL (CVs Lattes) para o formato JSON, facilitando a manipulação, análise e integração com outras ferramentas.

```
import os
import json
import xmltodict

def converter_xml_para_json(pasta_xml, pasta_saida):
    if not os.path.exists(pasta_saida):
        os.makedirs(pasta_saida)

    for arquivo in os.listdir(pasta_xml):
        if arquivo.endswith(".xml"):
            caminho_xml = os.path.join(pasta_xml, arquivo)
            nome_json = os.path.splitext(arquivo)[0] + ".json"
            caminho_json = os.path.join(pasta_saida, nome_json)

            with open(caminho_xml, "r", encoding="iso8859-1") as f_xml:
                try:
                    conteudo_xml = f_xml.read()
                    dados_dict = xmltodict.parse(conteudo_xml)
                    with open(caminho_json, "w", encoding="utf-8") as
f_json:
                        json.dump(dados_dict, f_json, indent=4,
ensure_ascii=False)
                    print(f"Convertido: {arquivo} -> {nome_json}")
                except Exception as e:
                    print(f"Erro ao converter {arquivo}: {e}")

pasta_dos_xmls = "xmls"
pasta_dos_jsons = "jsons"
converter_xml_para_json(pasta_dos_xmls, pasta_dos_jsons)
```

APÊNDICE B – *Script* para mesclagem de arquivos JSON.

Este *script* percorre uma pasta contendo múltiplos arquivos JSON e mescla todos os conteúdos em um único arquivo JSON. Ele é ideal para unificar saídas individuais geradas por outros *scripts* ou ferramentas.

```
import json
import glob
import os

pasta = 'jsons'

arquivos = glob.glob(os.path.join(pasta, '*.json'))

dados_combinados = []

for arquivo in arquivos:
    with open(arquivo, 'r', encoding='utf-8') as f:
        try:
            dados = json.load(f)
            if isinstance(dados, list):
                dados_combinados.extend(dados)
            else:
                dados_combinados.append(dados)
        except json.JSONDecodeError as e:
            print(f"Erro ao decodificar {arquivo}: {e}")

with open('jsons_mesclados.json', 'w', encoding='utf-8') as f:
    json.dump(dados_combinados, f, ensure_ascii=False, indent=2)

print(f"Mesclagem completa: {len(dados_combinados)} registros.")
```

APÊNDICE C – Extrator de metadados de produção científica.

Este *script* percorre uma pasta contendo arquivos JSON com produções científicas da PL e extrai metadados de artigos completos de revistas.

```
import json
import os

pasta_json = "jsons"
saida_json = "producao_cientifica.json"
producoes = []

for nome_arquivo in os.listdir(pasta_json):
    if not nome_arquivo.endswith(".json"):
        continue

    caminho = os.path.join(pasta_json, nome_arquivo)
    with open(caminho, "r", encoding="utf-8") as f:
        try:
            registros = json.load(f)
            if not isinstance(registros, dict):
                print(f"Estrutura inesperada em {nome_arquivo}")
                continue

            for id_producao, item in registros.items():
                if not isinstance(item, dict):
                    continue
                if not id_producao.startswith("ARTIGO-PUBLICADO/"):
                    continue

                dados_basicos = item.get("DADOS-BASICOS", {})
                if dados_basicos.get("@NATUREZA", "").strip().upper() !=
"COMPLETO":
                    continue

                autores = []
                for autor in item.get("AUTORES", []):
                    autores.append({
                        "nome": autor.get("@NOME-COMPLETO"),
                        "nro_id_cnpq": autor.get("@NRO-ID-CNPQ")
                    })

                producoes.append({
                    "id_producao": id_producao,
                    "titulo": dados_basicos.get("@TITULO"),
                    "ano": dados_basicos.get("@ANO"),
                    "idioma": dados_basicos.get("@IDIOMA"),
                    "autores": autores
                })

        except json.JSONDecodeError:
            print(f"Erro: JSON inválido em {nome_arquivo}")
        except Exception as e:
            print(f"Erro ao processar {nome_arquivo}: {e}")

with open(saida_json, "w", encoding="utf-8") as f_out:
    json.dump(producoes, f_out, ensure_ascii=False, indent=4)

print(f"Produções extraídas com sucesso: {saida_json}")
```

APÊNDICE D – Classe para serialização da produção científica e dos CVs Lattes

Classe responsável por realizar a leitura e serialização de arquivos JSON relacionados à produção científica e dados de CVs Lattes, convertendo-os em objetos Java. Também dá suporte à geração de classes Java a partir de esquemas JSON.

```
/**
 * Instrução: Importar na IDE como um projeto Maven.
*/

package utils;

import java.io.File;

public class UtilsMain {

      public static void main(String[] args) {

            String arquivo = "curriculos_consolidados.json";
//producao_cientifica.json
            GeneratorClassFromJson generator = new
GeneratorClassFromJson();
            generator.generateClasses("Base",
                                      "generated_curriculos_consolidados",
//generated_producao_cientifica
                                      "src/main/java",
                                      new File(arquivo));
      }
}
```

APÊNDICE E – Extrator de metadados de CVs Lattes.

Este *script* percorre uma pasta com arquivos JSON de CVs Lattes e extrai metadados de cada currículo, consolidando tudo em um único arquivo JSON.

```
import json
import os

credenciado = 'S' #variável alterada a mão de acordo com a finalidade
pasta_json = "jsons"
saida_json = "curriculos_consolidados.json"

curriculos_consolidados = []

if not os.path.exists(pasta_json):
    print(f"ERRO: A pasta '{pasta_json}' não foi encontrada.")
    exit()

print(f"Lendo arquivos de: {os.path.abspath(pasta_json)}")

for nome_arquivo in os.listdir(pasta_json):
    if nome_arquivo.endswith(".json"):
        caminho_arquivo = os.path.join(pasta_json, nome_arquivo)

        with open(caminho_arquivo, "r", encoding="utf-8") as f:
            try:
                conteudo = json.load(f)
                curriculo = conteudo.get("CURRICULO-VITAE", {})

                if not curriculo:
                    continue

                numero_identificador = curriculo.get("@NUMERO-
IDENTIFICADOR")

                dados_gerais = curriculo.get("DADOS-GERAIS", {})

                nome_completo = dados_gerais.get("@NOME-COMPLETO")
                pais_de_nascimento = dados_gerais.get("@PAIS-DE-
NASCIMENTO")
                uf_nascimento = dados_gerais.get("@UF-NASCIMENTO")
                cidade_nascimento = dados_gerais.get("@CIDADE-NASCIMENTO")

                endereco_data = dados_gerais.get("ENDERECO", {})
                if endereco_data is None: endereco_data = {}
                endereco_profissional = endereco_data.get("ENDERECO-
PROFISSIONAL")

                formacao_data = curriculo.get("FORMACAO-ACADEMICA-
TITULACAO")
                if not formacao_data:
                    formacao_data = dados_gerais.get("FORMACAO-ACADEMICA-
TITULACAO", {})
                if formacao_data is None: formacao_data = {}

                doutorado_raw = formacao_data.get("DOUTORADO")

                doc_conclusao = None
                doc_orientador = None
                doc_cod_inst = None
```

```
                doc_nome_inst = None
                alvo = None

                if isinstance(doutorado_raw, dict):
                if doutorado_raw.get("@STATUS-DO-CURSO") == "CONCLUIDO":
                        alvo = doutorado_raw
                elif isinstance(doutorado_raw, list):
                    concluidos = [d for d in doutorado_raw if
d.get("@STATUS-DO-CURSO") == "CONCLUIDO"]
                    if concluidos:
                        #ordena pelo ano (menor para o maior)
                        concluidos.sort(key=lambda x: int(x.get("@ANO-DE-
CONCLUSAO", "9999")))
                       alvo = concluidos[0]

                if alvo:
                    doc_conclusao = alvo.get("@ANO-DE-CONCLUSAO")
                    doc_orientador = alvo.get("@NUMERO-ID-ORIENTADOR")
                    doc_cod_inst = alvo.get("@CODIGO-INSTITUICAO")
                    doc_nome_inst = alvo.get("@NOME-INSTITUICAO")

                curriculos_consolidados.append({
                    "NUMERO-IDENTIFICADOR": numero_identificador,
                    "NOME-COMPLETO": nome_completo,
                    "PAIS-DE-NASCIMENTO": pais_de_nascimento,
                    "UF-NASCIMENTO": uf_nascimento,
                    "CIDADE-NASCIMENTO": cidade_nascimento,
                    "ENDERECO-PROFISSIONAL": endereco_profissional,

                    "DOUTORADO-ANO-CONCLUSAO": doc_conclusao,
                    "DOUTORADO-ORIENTADOR": doc_orientador,
                    "DOUTORADO-CODIGO-INSTITUICAO": doc_cod_inst,
                    "DOUTORADO-NOME-INSTITUICAO": doc_nome_inst,

                    "CREDENCIADO": credenciado
                })

            except json.JSONDecodeError:
                print(f"Ignorando {nome_arquivo}: JSON inválido.")
            except Exception as e:
                print(f"Erro no arquivo {nome_arquivo}: {e}")

with open(saida_json, "w", encoding="utf-8") as f_out:
    json.dump(curriculos_consolidados, f_out, ensure_ascii=False, indent=4)

print(f"Processo finalizado. {len(curriculos_consolidados)} currículos
processados.")
```

APÊNDICE F – Método de cálculo de distância.

Este método calcula a distância entre dois pontos geográficos na superfície da Terra, dados suas latitudes e longitudes, utilizando a fórmula de Haversine. A fórmula leva em consideração a curvatura do planeta para determinar a distância mais precisa entre os pontos, retornando o resultado em quilômetros. O método converte as coordenadas de graus para radianos, aplica a fórmula de Haversine para calcular a distância angular e, por fim, multiplica pelo raio médio da Terra (aproximadamente 6.371 km) para obter a distância.

```
/**
 * Calcula a distância entre dois pontos geográficos.
 */
public static double calculaDistancia(double lat1, double lon1, double
lat2, double lon2) {
    double R = 6371.0;

    double dLat = Math.toRadians(lat2 - lat1);
    double dLon = Math.toRadians(lon2 - lon1);

    lat1 = Math.toRadians(lat1);
    lat2 = Math.toRadians(lat2);

    double a = Math.sin(dLat / 2) * Math.sin(dLat / 2)
            + Math.cos(lat1) * Math.cos(lat2)
            * Math.sin(dLon / 2) * Math.sin(dLon / 2);

    double c = 2 * Math.atan2(Math.sqrt(a), Math.sqrt(1 - a));

    return R * c;
}
```

APÊNDICE G – Métodos de busca de latitude e longitude.

Estes métodos buscam a latitude e a longitude de um endereço fornecido, utilizando a API do OpenStreetMap (Nominatim). Eles utilizam requisição HTTP GET para o serviço, enviando o endereço codificado, e recebem resposta em formato JSON. Depois, analisam a resposta para extrair a latitude e longitude do primeiro resultado encontrado. Se o endereço for localizado, o método retorna a latitude e a longitude como um valor de ponto flutuante; caso contrário, retorna 0.0. É uma maneira prática de obter a coordenada geográfica de um endereço usando uma API de mapas.

```
package utils;

import java.io.BufferedReader;
import java.io.InputStreamReader;
import java.net.HttpURLConnection;
import java.net.URL;
import java.net.URLEncoder;
import java.nio.charset.StandardCharsets;
import java.util.HashMap;
import java.util.Map;
import org.json.JSONArray;
import org.json.JSONObject;

public class GeoUtils {

    private static final Map<String, String> ESTADOS = new HashMap<>();

    static {
        ESTADOS.put("AC", "Acre");
        ESTADOS.put("AL", "Alagoas");
        ESTADOS.put("AP", "Amapá");
        ESTADOS.put("AM", "Amazonas");
        ESTADOS.put("BA", "Bahia");
        ESTADOS.put("CE", "Ceará");
        ESTADOS.put("DF", "Distrito Federal");
        ESTADOS.put("ES", "Espírito Santo");
        ESTADOS.put("GO", "Goiás");
        ESTADOS.put("MA", "Maranhão");
        ESTADOS.put("MT", "Mato Grosso");
        ESTADOS.put("MS", "Mato Grosso do Sul");
        ESTADOS.put("MG", "Minas Gerais");
        ESTADOS.put("PA", "Pará");
        ESTADOS.put("PB", "Paraíba");
        ESTADOS.put("PR", "Paraná");
        ESTADOS.put("PE", "Pernambuco");
        ESTADOS.put("PI", "Piauí");
        ESTADOS.put("RJ", "Rio de Janeiro");
        ESTADOS.put("RN", "Rio Grande do Norte");
        ESTADOS.put("RS", "Rio Grande do Sul");
        ESTADOS.put("RO", "Rondônia");
        ESTADOS.put("RR", "Roraima");
        ESTADOS.put("SC", "Santa Catarina");
        ESTADOS.put("SP", "São Paulo");
        ESTADOS.put("SE", "Sergipe");
        ESTADOS.put("TO", "Tocantins");
    }

    /**
```

```
     * Retorna o nome da cidade acentuado e formatado (Ex: "São Carlos").
     */
    public static String retornaNomeCidadeOficial(String endereco) {
        JSONObject json = consultarApiNominatim(endereco);

        if (json != null && json.has("address")) {
            JSONObject address = json.getJSONObject("address");

            // O Nominatim pode retornar a cidade em campos diferentes
dependendo do tamanho
            if (address.has("city")) {
                return address.getString("city");
            } else if (address.has("town")) {
                return address.getString("town");
            } else if (address.has("municipality")) {
                return address.getString("municipality");
            } else if (address.has("village")) {
                return address.getString("village");
            } else if (address.has("hamlet")) {
                return address.getString("hamlet");
            }
        }
        return null;
    }

    /**
     * Retorna a Latitude.
     */
    public static Float retornaLatitude(String endereco) {
        JSONObject json = consultarApiNominatim(endereco);
        if (json != null) {
            return Float.valueOf(json.getString("lat"));
        }
        return 0f;
    }

    /**
     * Retorna a Longitude.
     */
    public static Float retornaLongitude(String endereco) {
        JSONObject json = consultarApiNominatim(endereco);
        if (json != null) {
            return Float.valueOf(json.getString("lon"));
        }
        return 0f;
    }

    /**
     * Método auxiliar privado para centralizar a chamada à API e tratar o
endereço.
     */
    private static JSONObject consultarApiNominatim(String
enderecoOriginal) {
        try {
            // 1. Tratamento do endereço
            String enderecoFormatado = tratarEndereco(enderecoOriginal);

            // 2. Construção da URL
            String query = URLEncoder.encode(enderecoFormatado,
StandardCharsets.UTF_8.toString());
```

```
            String urlString =
"https://nominatim.openstreetmap.org/search?format=json&addressdetails=1&li
mit=1&countrycodes=br&q=" + query;

            URL url = new URL(urlString);
            HttpURLConnection conn = (HttpURLConnection)
url.openConnection();
            conn.setRequestMethod("GET");
            conn.setRequestProperty("User-Agent", "Mozilla/5.0 (compatible;
SistemaAcademico/1.0)");

            if (conn.getResponseCode() != 200) {
                System.out.println("Erro HTTP: " + conn.getResponseCode());
                return null;
            }

            // Forçar UTF-8 na leitura para garantir acentuação correta
            BufferedReader in = new BufferedReader(
                new InputStreamReader(conn.getInputStream(),
StandardCharsets.UTF_8)
            );

            String inputLine;
            StringBuilder response = new StringBuilder();
            while ((inputLine = in.readLine()) != null) {
                response.append(inputLine);
            }
            in.close();

            JSONArray jsonArray = new JSONArray(response.toString());
            if (jsonArray.length() > 0) {
                return jsonArray.getJSONObject(0);
            }
        } catch (Exception e) {
            e.printStackTrace();
        }
        return null;
    }

    /**
     * Tenta identificar padrões como "Cidade/UF" e expandir para "Cidade,
Estado, Brazil".
     */
    private static String tratarEndereco(String endereco) {
        if (endereco == null) return "";
        if (endereco.contains("/")) {
            String[] partes = endereco.split("/");
            if (partes.length == 2) {
                String cidade = partes[0].trim();
                String sigla = partes[1].trim().toUpperCase();
                String nomeEstado = ESTADOS.getOrDefault(sigla, sigla);
                return cidade + ", " + nomeEstado + ", Brazil";
            }
        }
        if (!endereco.toLowerCase().contains("brazil") &&
!endereco.toLowerCase().contains("brasil")) {
             return endereco + ", Brazil";
        }

        return endereco;
    }
}
```

APÊNDICE H – Programa de mapeamento de coautorias científicas

Este programa mapeia relações de coautoria científica. A partir desses dados, gera um arquivo CSV com coordenadas geográficas para visualização cartográfica das colaborações, um arquivo de log para rastreabilidade do processamento e um arquivo auxiliar com IDs Lattes de coautores cujos currículos ainda não estão disponíveis, apoiando a coleta incremental de dados.

```
package main;

import com.google.gson.Gson;

import generated_curriculos_consolidados.Base;
import generated_curriculos_consolidados.Researcher;
import generated_curriculos_consolidados.EnderecoProfissional;

import generated_producao_cientifica.Publication;
import generated_producao_cientifica.Autore;

import json.JsonProcessor;
import utils.GeoUtils;

import java.io.BufferedWriter;
import java.io.File;
import java.io.FileWriter;
import java.io.IOException;

import java.text.Normalizer;
import java.util.*;
import java.util.concurrent.ConcurrentHashMap;

/**
 * Saídas:
 * - coordenadas.csv
 * - coautores.csv
 * - processamento.log
 */
public class Main {

    private static final String ARQUIVO_COORDENADAS = "coordenadas.csv";
    private static final String ARQUIVO_LOG = "processamento.log";
    private static final String ARQUIVO_COAUTORES = "coautores.csv";

    private static final String CIDADE_INSTITUICAO = "SAO CARLOS/SP";
    private static final String CODIGO_INSTITUICAO = "033500000006";
    private static final Integer ANO_LIMITE = 2024;

    // Cache thread-safe de coordenadas
    private static final Map<String, double[]> CACHE_GEO = new
ConcurrentHashMap<>();

    // Cache para nomes oficiais (com acento) para não chamar a API
repetidamente
    private static final Map<String, String> CACHE_NOMES_OFICIAIS = new
ConcurrentHashMap<>();

    public static void main(String[] args) {

        Gson gson = new Gson();
```

```
        JsonProcessor jsonProcessor = new JsonProcessor();

        // ============================
        // Leitura dos JSONs
        // ============================
        Base baseCC = gson.fromJson(
                jsonProcessor.readFromJsonFileToString(
                        new File("curriculos_consolidados.json")),
                Base.class
        );

        generated_producao_cientifica.Base basePC = gson.fromJson(
                jsonProcessor.readFromJsonFileToString(
                        new File("producao_cientifica.json")),
                generated_producao_cientifica.Base.class
        );

        // ============================
        // Indexação de pesquisadores por ID Lattes
        // ============================
        Map<String, Researcher> pesquisadoresPorId = new HashMap<>();
        for (Researcher r : baseCC.getResearchers()) {
            if (r.getNumeroIdentificador() != null &&
!r.getNumeroIdentificador().trim().isEmpty()) {
                pesquisadoresPorId.put(r.getNumeroIdentificador(), r);
            }
        }

        // ============================
        // Indexação de publicações por autor
        // ============================
        Map<String, List<Publication>> publicacoesPorAutor = new
HashMap<>();
        for (Publication pub : basePC.getPublications()) {
            for (Autore autor : pub.getAutores()) {
                String idAutor = autor.getNroIdCnpq();
                if (idAutor != null && !idAutor.trim().isEmpty()) {
                    publicacoesPorAutor
                            .computeIfAbsent(idAutor, k -> new
ArrayList<>())
                            .add(pub);
                }
            }
        }

        // ============================
        // Estrutura para deduplicação
        // ============================
        Set<String> coautoresUnicos = new HashSet<>();

        // ============================
        // Processamento principal
        // ============================
        try (
            BufferedWriter writerCoordenadas =
                    new BufferedWriter(new FileWriter(ARQUIVO_COORDENADAS,
true));
            BufferedWriter writerLog =
                    new BufferedWriter(new FileWriter(ARQUIVO_LOG, true));
            BufferedWriter writerCoautores =
                    new BufferedWriter(new FileWriter(ARQUIVO_COAUTORES,
true))
```

```
        ) {
            File arquivoCsv = new File(ARQUIVO_COORDENADAS);
            if (arquivoCsv.length() == 0) {

writerCoordenadas.write("id_producao;autor;coautor;ano;endogamica;latitude_
origem;longitude_origem;latitude_destino;longitude_destino;distancia_km;cid
ade_instituicao_destino;nome_instituicao_destino;codigo_instituicao_destino
;doutorado_ano_conclusao;doutorado_nome_instituicao;doutorado_codigo_instit
uicao");
                writerCoordenadas.newLine();
            }

            for (Researcher pesquisador : baseCC.getResearchers()) {
                if (!pesquisador.getCredenciado().equals("S")) {
                    continue;
                }

                String idPesquisador =
pesquisador.getNumeroIdentificador();
                if (idPesquisador == null ||
idPesquisador.trim().isEmpty()) continue;

                EnderecoProfissional endPesquisador =
pesquisador.getEnderecoProfissional();
                if (endPesquisador == null ||
endPesquisador.getCidade().isEmpty() || endPesquisador.getUf().isEmpty()) {
                    continue;
                }

                String cidadeLimpa =
padronizarNomeCidade(endPesquisador.getCidade());
                String ufLimpa =
endPesquisador.getUf().trim().toUpperCase();

                // Chave padronizada para lógica (SAO CARLOS/SP)
                String cidadeInstituicaoPesquisadorKey = cidadeLimpa + "/"
+ ufLimpa;

                // Nome para exibição (São Carlos/SP)
                String cidadeInstituicaoPesquisadorDisplay =
obterNomeCidadeOficial(cidadeInstituicaoPesquisadorKey) + "/" + ufLimpa;

                String nomeInstituicaoPesquisador =
endPesquisador.getNomeInstituicaoEmpresa();

                // Verifica usando a chave padronizada
                if
(!cidadeInstituicaoPesquisadorKey.equals(CIDADE_INSTITUICAO) ||
!CODIGO_INSTITUICAO.equals(endPesquisador.getCodigoInstituicaoEmpresa())) {
                    continue;
                }

                List<Publication> publicacoes =
publicacoesPorAutor.get(idPesquisador);
                if (publicacoes == null) continue;

                double[] geoOrigem = geo(cidadeInstituicaoPesquisadorKey);

                for (Publication pub : publicacoes) {
                    String idProducao = pub.getIdProducao();

                    Integer anoPublicacao = parseIntSafe(pub.getAno());
```

```
                    if (anoPublicacao == null || anoPublicacao >
ANO_LIMITE) continue;

                    for (Autore a : pub.getAutores()) {
                        String endogamica = "N";

                        String idCoautor = a.getNroIdCnpq();

                        if (idCoautor == null || idCoautor.trim().isEmpty()
|| idCoautor.equals(idPesquisador)) {
                            continue;
                        }

                        if (!idCoautor.trim().isEmpty() &&
!pesquisadoresPorId.containsKey(idCoautor) &&
coautoresUnicos.add(idCoautor)) {
                            writerCoautores.write(idCoautor.trim());
                            writerCoautores.newLine();
                        }

                        Researcher coautor =
pesquisadoresPorId.get(idCoautor);

                        if (coautor == null) continue;

                        Integer anoDoutorado =
parseIntSafe(coautor.getDoutoradoAnoConclusao());
                        if (anoDoutorado != null && anoPublicacao >
anoDoutorado && idPesquisador.equals(coautor.getDoutoradoOrientador())) {
                            endogamica = "S";
                        }

                        EnderecoProfissional endCoautor =
coautor.getEnderecoProfissional();
                        if (endCoautor == null ||
endCoautor.getCidade().isEmpty() || endCoautor.getUf().isEmpty()) {
                            continue;
                        }

                        String cidadeCoautorLimpa =
padronizarNomeCidade(endCoautor.getCidade());
                        String ufCoautorLimpa =
endCoautor.getUf().trim().toUpperCase();

                        // Chave padronizada (ex: ARARAQUARA/SP)
                        String cidadeInstituicaoCoautorKey =
cidadeCoautorLimpa + "/" + ufCoautorLimpa;

                        // Nome para exibição (ex: Araraquara/SP)
                        String cidadeInstituicaoCoautorDisplay =
obterNomeCidadeOficial(cidadeInstituicaoCoautorKey) + "/" + ufCoautorLimpa;

                        String nomeInstituicaoCoautor =
endCoautor.getNomeInstituicaoEmpresa();
                        String codigoInstituicaoCoautor =
endCoautor.getCodigoInstituicaoEmpresa();

                        // Busca coordenadas usando a chave padronizada
                        double[] geoDestino =
geo(cidadeInstituicaoCoautorKey);

                        double distanciaKm = GeoUtils.calculaDistancia(
```

```
                                geoOrigem[0], geoOrigem[1],
                                geoDestino[0], geoDestino[1]
                        );

                        // Escrita no CSV usando o nome com acento
                        writerCoordenadas.write(
                                String.format(
                                        Locale.US,

"%s;%s;%s;%s;%s;%.6f;%.6f;%.6f;%.6f;%.2f;%s;%s;%s;%s;%s;%s%n",
                                        idProducao,
                                        idPesquisador,
                                        idCoautor,
                                        anoPublicacao,
                                        endogamica,
                                        geoOrigem[0],
                                        geoOrigem[1],
                                        geoDestino[0],
                                        geoDestino[1],
                                        distanciaKm,
                                        cidadeInstituicaoCoautorDisplay,
                                        nomeInstituicaoCoautor.replace(";",
",","),
                                        codigoInstituicaoCoautor,
                                        (coautor.getDoutoradoAnoConclusao()
!= null ? coautor.getDoutoradoAnoConclusao() : ""),

(coautor.getDoutoradoNomeInstituicao() != null ?
coautor.getDoutoradoNomeInstituicao().replace(";", ",") : ""),

(coautor.getDoutoradoCodigoInstituicao() != null ?
coautor.getDoutoradoCodigoInstituicao() : "")
                                        )
                        );

                        StringBuilder log = new StringBuilder(512);

                        log.append("ID_Producao: ")
                        .append(pub.getIdProducao())
                        .append(" | ")
                        .append("Endogâmica: ")
                        .append(endogamica)
                        .append("\n");

                        log.append("Autor: ")
                           .append(pesquisador.getNomeCompleto())
                           .append(" (ID Lattes: ")
                           .append(idPesquisador)
                           .append(" | ")
                           .append(cidadeInstituicaoPesquisadorDisplay)
                           .append(" | ")
                           .append(nomeInstituicaoPesquisador)
                           .append(")\n");

                        log.append("--> Coautor: ")
                           .append(coautor.getNomeCompleto())
                           .append(" (ID Lattes: ")
                           .append(idCoautor)
                           .append(" | ")
                           .append(cidadeInstituicaoCoautorDisplay)
                           .append(" | ")
                           .append(nomeInstituicaoCoautor)
```

```
                            .append(")\n");

                        log.append("Título: ")
                            .append(pub.getTitulo())
                            .append("\n");

                        log.append("Ano publicação: ")
                            .append(anoPublicacao)
                            .append(" | Ano doutorado: ")
                            .append(anoDoutorado)
                            .append("\n");

                        log.append(
                            "-----------------------------------------------
---\n");

                        writerLog.write(log.toString());
                    }
                }
            }
        } catch (IOException e) {
            e.printStackTrace();
        }

        System.out.println("--> Fim do processamento.");
    }

    // ============================
    // Métodos Auxiliares
    // ============================

    private static String padronizarNomeCidade(String cidade) {
        if (cidade == null) return "";
        return Normalizer.normalize(cidade, Normalizer.Form.NFKD)
                .replaceAll("\\p{M}", "")
                .trim()
                .toUpperCase(Locale.ROOT);
    }

    private static String obterNomeCidadeOficial(String
cidadeUfPadronizada) {
        return CACHE_NOMES_OFICIAIS.computeIfAbsent(cidadeUfPadronizada, k
-> {
            // Tenta buscar o nome na API
            String nomeOficial = GeoUtils.retornaNomeCidadeOficial(k);

            // Pega o nome bruto (seja da API ou do split)
            String nomeBruto = (nomeOficial != null) ? nomeOficial :
k.split("/")[0];

            // Aplica a formatação "Title Case" antes de salvar no cache
            return formatarTitleCase(nomeBruto);
        });
    }

    private static String formatarTitleCase(String texto) {
        if (texto == null || texto.isEmpty()) {
            return texto;
        }

        // Converte tudo para minúsculo e separa por espaços
        String[] palavras = texto.toLowerCase().split("\\s+");
```

```
        StringBuilder resultado = new StringBuilder();

        for (String palavra : palavras) {
            // Lista de preposições que devem continuar minúsculas
            if (palavra.matches("^(e|de|da|do|das|dos)$")) {
                resultado.append(palavra);
            } else if (palavra.length() > 0) {
                // Capitaliza a primeira letra das outras palavras
                resultado.append(Character.toUpperCase(palavra.charAt(0)))
                         .append(palavra.substring(1));
            }
            resultado.append(" ");
        }

        return resultado.toString().trim();
    }


    private static double[] geo(String cidadeUf) {
        return CACHE_GEO.computeIfAbsent(
                cidadeUf,
                k -> new double[]{
                        GeoUtils.retornaLatitude(k),
                        GeoUtils.retornaLongitude(k)
                }
        );
    }

    private static Integer parseIntSafe(String valor) {
        if (valor == null) return null;
        try {
            String v = valor.trim();
            if (v.isEmpty()) return null;
            return Integer.parseInt(v);
        } catch (Exception e) {
            return null;
        }
    }
}
```

APÊNDICE I – *Script* para contagem de ocorrências de latitude e longitude.

Este *script* lê o arquivo CSV com coordenadas, agrupa os dados pelas colunas de latitude e longitude de origem e destino, conta quantas ocorrências existem para cada combinação, e salva esse resultado em um novo arquivo CSV.

```
import pandas as pd

endogamica = 'S' #variável alterada a mão de acordo com a finalidade

caminho_entrada = 'coordenadas.csv'
caminho_saida = 'ocorrencias.csv'

df = pd.read_csv(caminho_entrada, sep=';')

df_filtrado = df[df['endogamica'].astype(str).str.strip().str.upper() ==
endogamica].copy()

if len(df_filtrado) > 0:
    resultado = df_filtrado.groupby([
        'latitude_origem'
        ,'longitude_origem'
        ,'latitude_destino'
        ,'longitude_destino'
        ,'cidade_instituicao_destino'
    ]).size().reset_index(name='quantidade')

    resultado.to_csv(caminho_saida, sep=';', index=False)

    print("Arquivo gerado com sucesso:", caminho_saida)
    print(resultado.head())
else:
    print("Nenhum registro encontrado para o filtro aplicado.")
```

APÊNDICE J – *Script* de geração de mapa interativo.

Este *script* gera um mapa interativo, a partir de um arquivo CSV com dados de coordenadas e ocorrências. Ele plota pontos de origem e destino, por meio de elementos como linhas e setas.

```
import folium
import pandas as pd
from folium.plugins import PolyLineTextPath
from folium import Element

def cor_e_tamanho_flecha(quantidade):
    if quantidade > 100:
        return 'red', '22px'
    elif quantidade > 50:
        return 'orange', '18px'
    elif quantidade > 20:
        return 'blue', '14px'
    else:
        return 'gray', '12px'

def formata_milhar(valor):
    return f"{valor:,}".replace(",", ".")

def criar_mapa_ocorrencias(caminho_csv: str, caminho_saida_html: str =
"mapa.html") -> None:
    df = pd.read_csv(caminho_csv, sep=';')

    df = df.dropna(subset=[
        'latitude_origem', 'longitude_origem',
        'latitude_destino', 'longitude_destino',
        'quantidade'
    ])

    mapa = folium.Map(
        location=[df['latitude_origem'].mean(),
df['longitude_origem'].mean()],
        zoom_start=5
    )

    marcadores_adicionados = set()

    for _, row in df.iterrows():
        lat_origem = row['latitude_origem']
        lon_origem = row['longitude_origem']
        lat_destino = row['latitude_destino']
        lon_destino = row['longitude_destino']
        origem = (lat_origem, lon_origem)
        destino = (lat_destino, lon_destino)
        cidade_destino = row['cidade_instituicao_destino']
        quantidade = int(row['quantidade'])

        if origem == destino:
            folium.Marker(
                location=origem,
                popup=f"{formata_milhar(quantidade)} ocorrências internas",
                tooltip=f"{formata_milhar(quantidade)}          ocorrências
internas",
                icon=folium.Icon(color='red', icon='home', prefix='fa')
            ).add_to(mapa)
```

```
            marcadores_adicionados.add(origem)
            continue

        if origem not in marcadores_adicionados:
            folium.Marker(
                location=origem,
                icon=folium.Icon(color='red', icon='home', prefix='fa')
            ).add_to(mapa)
            marcadores_adicionados.add(origem)

        if destino not in marcadores_adicionados:
            folium.Marker(
                location=destino,
                popup=f"{cidade_destino}:      {formata_milhar(quantidade)}
ocorrências",
                tooltip=f"{cidade_destino}:    {formata_milhar(quantidade)}
ocorrências",
                icon=folium.Icon(color='red', icon='location')
            ).add_to(mapa)
            marcadores_adicionados.add(destino)

        linha = folium.PolyLine(
            locations=[origem, destino],
            color='black',
            weight=3,  # Espessura fixa
            opacity=0.6,
            tooltip=f'{quantidade} ocorrências'
        ).add_to(mapa)

        cor, tamanho = cor_e_tamanho_flecha(quantidade)
        PolyLineTextPath(
            linha,
            '➤',
            repeat=True,
            offset=7,
            attributes={
                'fill': cor,
                'font-weight': 'bold',
                'font-size': tamanho
            }
        ).add_to(mapa)

    legenda_html = '''
    <div style="
        position: fixed;
        bottom: 50px;
        right: 50px;
        z-index: 9999;
        background-color: white;
        padding: 12px;
        border: 2px solid grey;
        border-radius: 10px;
        box-shadow: 2px 2px 5px rgba(0,0,0,0.3);
        font-size: 14px;
        line-height: 1.6;
        font-family: Arial, sans-serif;
    ">
        <strong>Legenda:</strong><br><br>
        <u>Cor das setas (quantidade):</u><br>
        <span style="color: red;">&#10148;</span> Mais de 100<br>
        <span style="color: orange;">&#10148;</span> 51 a 100<br>
```

```
        <span style="color: blue;">&#10148;</span> 21 a 50<br>
        <span style="color: gray;">&#10148;</span> 20 ou menos
    </div>
    '''
    mapa.get_root().html.add_child(Element(legenda_html))

    mapa.save(caminho_saida_html)
    print(f"Mapa gerado com sucesso em: {caminho_saida_html}")

if __name__ == "__main__":
    criar_mapa_ocorrencias("ocorrencias.csv")
```

APÊNDICE K – *Script* para visualização da porcentagem de endogamia das coautorias

O *script* contabiliza a frequência de registros classificados como endogâmicos e não endogâmicos. Em seguida, gera um gráfico de pizza para visualizar a proporção percentual entre as coautorias que possuem ou não a presença de endogamia.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

sns.set_style("white")

arquivo = 'coordenadas.csv'

try:
    df = pd.read_csv(arquivo, sep=';', encoding='utf-8', engine='python')
except UnicodeDecodeError:
    df = pd.read_csv(arquivo, sep=';', encoding='latin1', engine='python')

df['endogamica'] = df['endogamica'].astype(str).str.strip().str.upper()

df['endogamica'] = df['endogamica'].replace({'S': 'Endogâmicas', 'N': 'Não
endogâmicas'})

contagem = df['endogamica'].value_counts()

plt.figure(figsize=(8, 8), facecolor='white')

cores = ['#66b3ff', '#ff9999']

plt.pie(
    contagem,
    labels=contagem.index,
    autopct='%1.0f%%',
    startangle=140,
    colors=cores,
    textprops={'fontsize': 14}
)

plt.title('Proporção de tipos de coautorias', fontsize=16)
plt.axis('equal')

plt.savefig('endogamia_tipos_coautorias.png', dpi=300, bbox_inches='tight',
facecolor='white')
plt.show()
```

APÊNDICE L – *Script* para visualização da porcentagem de endogamia das publicações

O *script* agrupa os registros de coautorias por publicação para identificar quais trabalhos possuem incidência de endogamia. Em seguida, gera um gráfico de pizza para visualizar a proporção percentual entre as publicações com e sem endogamia.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

sns.set_style("white")

try:
    df = pd.read_csv('coordenadas.csv', sep=';', encoding='utf-8',
engine='python')
except UnicodeDecodeError:
    df = pd.read_csv('coordenadas.csv', sep=';', encoding='latin1',
engine='python')

df['is_endogamic'] =
df['endogamica'].astype(str).str.strip().str.upper().apply(lambda x: 1 if x
== 'S' else 0)

publicacoes_agrupadas =
df.groupby('id_producao')['is_endogamic'].max().reset_index()

publicacoes_agrupadas['status'] =
publicacoes_agrupadas['is_endogamic'].apply(
    lambda x: 'Com endogamia' if x == 1 else 'Sem endogamia'
)

contagem = publicacoes_agrupadas['status'].value_counts()

plt.figure(figsize=(10, 8), facecolor='white')

cores = ['#66b3ff', '#ff9999']

plt.pie(
    contagem,
    labels=contagem.index,
    autopct='%1.0f%%',
    startangle=140,
    colors=cores,
    textprops={'fontsize': 14}
)

plt.title('Proporção de tipos de publicações', fontsize=16, pad=20)
plt.axis('equal')

plt.tight_layout()
plt.savefig('endogamia_tipos_publicacoes.png', dpi=300,
bbox_inches='tight', facecolor='white')
plt.show()
```

APÊNDICE M – *Script* para visualização da endogamia das publicações

O *script* calcula a taxa de coautoria endogâmica de cada publicação. Em seguida, gera um histograma para visualizar quantas publicações se encontram em cada faixa de endogamia.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

sns.set_style("white")

try:
    df = pd.read_csv('coordenadas.csv', sep=';', encoding='utf-8')
except UnicodeDecodeError:
    df = pd.read_csv('coordenadas.csv', sep=';', encoding='latin1')

df['is_endogamic'] =
df['endogamica'].astype(str).str.strip().str.upper().apply(lambda x: 1 if x
== 'S' else 0)

publication_stats =
df.groupby('id_producao')['is_endogamic'].mean().reset_index()
publication_stats['pct_endogamica'] = publication_stats['is_endogamic'] *
100

largura_bin = 5
bins_centralizados = np.arange(-2.5, 105, largura_bin)

plt.figure(figsize=(12, 6))

sns.histplot(
    data=publication_stats,
    x='pct_endogamica',
    bins=bins_centralizados,
    kde=True,
    color='grey',
    edgecolor='black',
    line_kws={'linewidth': 2}
)

plt.title('Distribuição da porcentagem de endogamia de publicações',
fontsize=14)
plt.xlabel('Porcentagem de coautorias endogâmicas (%)', fontsize=12)
plt.ylabel('Quantidade de publicações', fontsize=12)

plt.xticks(np.arange(0, 101, 10))
plt.xlim(-5, 105)

plt.tight_layout()
plt.savefig('endogamia_publicacao.png', dpi=300, bbox_inches='tight',
facecolor='white')
plt.show()
```

APÊNDICE N – *Script* para visualização da endogamia dos pesquisadores

O *script* calcula a taxa de coautoria endogâmica de cada pesquisador. Em seguida, gera um histograma para visualizar quantos pesquisadores se encontram em cada faixa de endogamia.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

sns.set_style("white")

try:
    df = pd.read_csv('coordenadas.csv', sep=';', encoding='utf-8')
except UnicodeDecodeError:
    df = pd.read_csv('coordenadas.csv', sep=';', encoding='latin1')

df['is_endogamic'] =
df['endogamica'].astype(str).str.strip().str.upper().apply(lambda x: 1 if x
== 'S' else 0)

author_stats = df.groupby('autor')['is_endogamic'].mean().reset_index()
author_stats['pct_endogamica'] = author_stats['is_endogamic'] * 100

largura_bin = 5
bins_centralizados = np.arange(-2.5, 105, largura_bin)

plt.figure(figsize=(12, 6))

sns.histplot(
    data=author_stats,
    x='pct_endogamica',
    bins=bins_centralizados,
    kde=True,
    color='grey',
    edgecolor='black',
    line_kws={'linewidth': 2}
)

plt.title('Distribuição da porcentagem de endogamia de pesquisadores',
fontsize=14)
plt.xlabel('Porcentagem de coautorias endogâmicas (%)', fontsize=12)
plt.ylabel('Quantidade de pesquisadores', fontsize=12)

plt.xticks(np.arange(0, 101, 10))
plt.xlim(-5, 105)

plt.tight_layout()
plt.savefig('endogamia_pesquisador.png', dpi=300, bbox_inches='tight',
facecolor='white')
plt.show()
```

APÊNDICE O – *Script* para visualização da endogamia das instituições

Este *script* possibilita a análise da relação entre a quantidade de publicações e a porcentagem de coautorias endogâmicas das instituições.

```
import pandas as pd
import altair as alt

arquivo_csv = "coordenadas.csv"
saida = "instituicao_publicacao_coautoria_endogamica.html"

csv_params = {'sep': ";", 'dtype': {'codigo_instituicao_destino': str},
'on_bad_lines': 'warn'}

try:
    df = pd.read_csv(arquivo_csv, encoding="utf-8", **csv_params)
except:
    df = pd.read_csv(arquivo_csv, encoding="latin1", engine='python',
on_bad_lines='skip', sep=";", dtype={'codigo_instituicao_destino': str})

df = df.dropna(subset=['codigo_instituicao_destino'])

df["endogamica"] = df["endogamica"].astype(str).str.strip().str.upper()
df["is_endogamica"] = (df["endogamica"] == 'S').astype(int)

df_inst = df.groupby(['codigo_instituicao_destino',
'nome_instituicao_destino', 'cidade_instituicao_destino']).agg(
    qtd_publicacoes=('id_producao', 'nunique'),
    total_coautorias=('id_producao', 'count'),
    qtd_endogamicas=('is_endogamica', 'sum')
).reset_index()

df_inst['pct_endogamia'] = df_inst['qtd_endogamicas'] /
df_inst['total_coautorias']

max_x = df_inst['qtd_publicacoes'].max()
dominio_x = [0.8, max_x * 1.5]

dominios_tamanho = [10, 100, 1000]
tamanhos_area = [200, 500, 1000, 2000]

pt_br_locale = {"decimal": ",", "thousands": ".", "grouping": [3],
"currency": ["R$", ""]}

chart = alt.Chart(df_inst).mark_circle(
    opacity=0.7,
    stroke='black',
    strokeWidth=1,
    color='#2171b5'
).encode(
    tooltip=[
        alt.Tooltip('nome_instituicao_destino', title='Instituição:'),
        alt.Tooltip('cidade_instituicao_destino', title='Cidade:'),
        alt.Tooltip('qtd_publicacoes', title='Qtde de publicações:',
format=',.0d'),
        alt.Tooltip('total_coautorias', title='Qtde de coautorias:',
format=',.0d'),
        alt.Tooltip('pct_endogamia', title='Nível de endogamia:',
format='.0%')
    ],
```

```
    x=alt.X('qtd_publicacoes:Q',
            title='Quantidade de publicações',
            scale=alt.Scale(type='log', domain=dominio_x)
    ),

    y=alt.Y('pct_endogamia:Q',
            title='Porcentagem de coautorias endogâmicas',
            scale=alt.Scale(domain=[-0.05, 1.05], nice=False),
            axis=alt.Axis(format='%', values=[0, 0.2, 0.4, 0.6, 0.8, 1.0])
    ),

    size=alt.Size('total_coautorias:Q',
                  title='Qtde de coautorias',
                  scale=alt.Scale(
                      type='threshold',
                      domain=dominios_tamanho,
                      range=tamanhos_area
                  ),
                  legend=alt.Legend(
                      title="Qtde de coautorias",
                      values=[1, 10, 100, 1000],
                      format=",.0d",
                      symbolFillColor='#2171b5',
                      symbolStrokeColor='black',
                      symbolStrokeWidth=1
                  )
    )

).properties(
    title='Endogamia por Instituição (Publicação vs Coautoria endogâmica)',
    width=900,
    height=600
).configure(
    locale={"number": pt_br_locale}
).interactive()

chart.save(saida)
print(f"\nVisualização salva: {saida}")
```

APÊNDICE P – *Script* para visualização da endogamia das instituições por distância

Este *script* possibilita a análise da relação entre a distância e a porcentagem de coautorias endogâmicas das instituições. A distância tem como referência de origem a UFSCar – Campus São Carlos.

```
import pandas as pd
import altair as alt

arquivo_csv = "coordenadas.csv"
saida_html = "instituicao_distancia_coautoria_endogamica.html"

csv_params = {'sep': ";", 'dtype': {'codigo_instituicao_destino': str},
'on_bad_lines': 'warn'}

try:
    df = pd.read_csv(arquivo_csv, encoding="utf-8", **csv_params)
except:
    df = pd.read_csv(arquivo_csv, encoding="latin1", engine='python',
on_bad_lines='skip', sep=";", dtype={'codigo_instituicao_destino': str})

df = df.dropna(subset=['codigo_instituicao_destino'])

df["endogamica"] = df["endogamica"].astype(str).str.strip().str.upper()
df["is_endogamica"] = (df["endogamica"] == 'S').astype(int)

df_inst = df.groupby(['codigo_instituicao_destino',
'nome_instituicao_destino', 'cidade_instituicao_destino']).agg(
    distancia_media=('distancia_km', 'mean'),
    total_coautorias=('id_producao', 'count'),
    qtd_endogamicas=('is_endogamica', 'sum')
).reset_index()

df_inst['pct_endogamia'] = df_inst['qtd_endogamicas'] /
df_inst['total_coautorias']

LIMITE_DISTANCIA = 2500
PASSO = 250

POSICAO_MAIS = LIMITE_DISTANCIA + PASSO

def agrupar_distancia(d):
    if d > LIMITE_DISTANCIA:
        return POSICAO_MAIS
    return d

df_inst['distancia_plot'] =
df_inst['distancia_media'].apply(agrupar_distancia)

ticks_x = list(range(0, LIMITE_DISTANCIA + 1, PASSO)) + [POSICAO_MAIS]

dominio_x = [-150, POSICAO_MAIS + 150]

pt_br_locale = {"decimal": ",", "thousands": ".", "grouping": [3],
"currency": ["R$", ""]}
dominios_tamanho = [10, 100, 1000]
tamanhos_area = [200, 500, 1000, 2000]

chart = alt.Chart(df_inst).mark_circle(
    opacity=0.7,
```

```
    stroke='black',
    strokeWidth=1,
    color='#2171b5'
).encode(
    tooltip=[
        alt.Tooltip('nome_instituicao_destino', title='Instituição:'),
        alt.Tooltip('cidade_instituicao_destino', title='Cidade:'),
        alt.Tooltip('distancia_media', title='Distância (km):',
format=",.0d"),
        alt.Tooltip('total_coautorias', title='Qtde de coautorias:',
format=",.0d"),
        alt.Tooltip('pct_endogamia', title='Nível de endogamia:',
format='.0%')
    ],

    x=alt.X('distancia_plot:Q',
            title='Distância (km) da UFSCar – Campus São Carlos',
            scale=alt.Scale(domain=dominio_x),
            axis=alt.Axis(
                values=ticks_x,
                labelExpr=f"datum.value == {POSICAO_MAIS} ? '2.500+' :
format(datum.value, ',.0d')"
            )
    ),

    y=alt.Y('pct_endogamia:Q',
            title='Porcentagem de coautorias endogâmicas',
            scale=alt.Scale(domain=[-0.1, 1.1], nice=False),
            axis=alt.Axis(format='%', values=[0, 0.2, 0.4, 0.6, 0.8, 1.0])
    ),

    size=alt.Size('total_coautorias:Q',
                  title='Qtde de coautorias',
                  scale=alt.Scale(type='threshold',
domain=dominios_tamanho, range=tamanhos_area),
                  legend=alt.Legend(
                      title="Qtde de coautorias",
                      values=[1, 10, 100, 1000],
                      format=",.0d",
                      symbolFillColor='#2171b5',
                      symbolStrokeColor='black',
                      symbolStrokeWidth=1
                  )
    )

).properties(
    title='Endogamia por Instituição (Distância vs Coautoria endogâmica)',
    width=1000,
    height=600
).configure(
    locale={"number": pt_br_locale}
).interactive()

chart.save(saida_html)
print(f"\nVisualização salva com sucesso em: {saida_html}")
```

APÊNDICE Q – *Script* para visualização da coautoria das instituições

Este *script* possibilita a análise da relação entre a distância e a quantidade de coautorias das instituições. A distância tem como referência de origem a UFSCar – Campus São Carlos.

```
import pandas as pd
import altair as alt

arquivo_csv = "coordenadas.csv"
saida_html = "instituicao_distancia_coautoria.html"

csv_params = {'sep': ";", 'dtype': {'codigo_instituicao_destino': str},
'on_bad_lines': 'warn'}

try:
    df = pd.read_csv(arquivo_csv, encoding="utf-8", **csv_params)
except:
    df = pd.read_csv(arquivo_csv, encoding="latin1", engine='python',
on_bad_lines='skip', sep=";", dtype={'codigo_instituicao_destino': str})

df = df.dropna(subset=['codigo_instituicao_destino'])

df["endogamica"] = df["endogamica"].astype(str).str.strip().str.upper()
df["is_endogamica"] = (df["endogamica"] == 'S').astype(int)

df_inst = df.groupby(['codigo_instituicao_destino',
'nome_instituicao_destino', 'cidade_instituicao_destino']).agg(
    distancia_media=('distancia_km', 'mean'),
    total_coautorias=('id_producao', 'count'),
    qtd_endogamicas=('is_endogamica', 'sum')
).reset_index()

df_inst['pct_endogamia'] = df_inst['qtd_endogamicas'] /
df_inst['total_coautorias']

LIMITE_X = 2500
PASSO_X = 250
POSICAO_MAIS_X = LIMITE_X + PASSO_X

def agrupar_distancia(d):
    if d > LIMITE_X:
        return POSICAO_MAIS_X
    return d

df_inst['distancia_plot'] =
df_inst['distancia_media'].apply(agrupar_distancia)
ticks_x = list(range(0, LIMITE_X + 1, PASSO_X)) + [POSICAO_MAIS_X]
dominio_x = [-150, POSICAO_MAIS_X + 150]

LIMITE_Y = 2500
PASSO_Y = 250
POSICAO_MAIS_Y = LIMITE_Y + PASSO_Y

def agrupar_quantidade(q):
    if q > LIMITE_Y:
        return POSICAO_MAIS_Y
    return q
```

```
df_inst['total_coautorias_plot'] =
df_inst['total_coautorias'].apply(agrupar_quantidade)

ticks_y = list(range(0, LIMITE_Y + 1, PASSO_Y)) + [POSICAO_MAIS_Y]
dominio_y = [-100, POSICAO_MAIS_Y + 150]

pt_br_locale = {"decimal": ",", "thousands": ".", "grouping": [3],
"currency": ["R$", ""]}

chart = alt.Chart(df_inst).mark_circle(
    opacity=0.7,
    stroke='black',
    strokeWidth=1,
    size=400
).encode(
    tooltip=[
        alt.Tooltip('nome_instituicao_destino', title='Instituição:'),
        alt.Tooltip('cidade_instituicao_destino', title='Cidade:'),
        alt.Tooltip('distancia_media', title='Distância (km):',
format=",.0d"),
        alt.Tooltip('total_coautorias', title='Qtde de coautorias:',
format=",.0d"),
        alt.Tooltip('pct_endogamia', title='Nível de endogamia:',
format='.0%')
    ],

    x=alt.X('distancia_plot:Q',
            title='Distância (km) da UFSCar – Campus São Carlos',
            scale=alt.Scale(domain=dominio_x),
            axis=alt.Axis(
                values=ticks_x,
                labelExpr=f"datum.value == {POSICAO_MAIS_X} ? '2.500+' :
format(datum.value, ',.0d')"
            )
    ),

    y=alt.Y('total_coautorias_plot:Q',
            title='Quantidade de coautorias',
            scale=alt.Scale(domain=dominio_y),
            axis=alt.Axis(
                values=ticks_y,
                labelExpr=f"datum.value == {POSICAO_MAIS_Y} ? '2.500+' :
format(datum.value, ',.0d')"
            )
    ),

    color=alt.Color('pct_endogamia:Q',
                    title='Endogamia',
                    scale=alt.Scale(scheme='reds', domain=[0, 1]),
                    legend=alt.Legend(format=".0%")
    )
).properties(
    title='Coautoria por Instituição (Distância vs Coautoria)',
    width=1000,
    height=1200
).configure(
    locale={"number": pt_br_locale}
).interactive()

chart.save(saida_html)
print(f"\nVisualização salva com sucesso em: {saida_html}")
```

APÊNDICE R – *Script* para visualização do total de publicações endogâmicas das instituições

Este *script* possibilita a análise da relação entre a distância e a quantidade de publicações endogâmicas das instituições. A distância tem como referência de origem a UFSCar – Campus São Carlos.

```
import pandas as pd
import altair as alt

arquivo_csv = "coordenadas.csv"
saida_html = "instituicao_distancia_com_endogamia.html"
csv_params = {'sep': ";", 'dtype': {'codigo_instituicao_destino': str},
'on_bad_lines': 'warn'}

try:
    df = pd.read_csv(arquivo_csv, encoding="utf-8", **csv_params)
except:
    df = pd.read_csv(arquivo_csv, encoding="latin1", engine='python',
on_bad_lines='skip', sep=";", dtype={'codigo_instituicao_destino': str})

df = df.dropna(subset=['codigo_instituicao_destino'])
df["endogamica"] = df["endogamica"].astype(str).str.strip().str.upper()
ids_com_endogamia = df.loc[df['endogamica'] == 'S', 'id_producao'].unique()
df = df[df['id_producao'].isin(ids_com_endogamia)].copy()
df["is_endogamica"] = (df["endogamica"] == 'S').astype(int)
df_inst = df.groupby(['codigo_instituicao_destino',
'nome_instituicao_destino', 'cidade_instituicao_destino']).agg(
    distancia_media=('distancia_km', 'mean'),
    qtd_publicacoes=('id_producao', 'nunique'),
    total_registros=('id_producao', 'count'),
    qtd_endogamicas=('is_endogamica', 'sum')
).reset_index()
df_inst['pct_endogamia'] = df_inst['qtd_endogamicas'] /
df_inst['total_registros']

LIMITE_X = 2500
PASSO_X = 250
POSICAO_MAIS_X = LIMITE_X + PASSO_X

def agrupar_distancia(d):
    if d > LIMITE_X:
        return POSICAO_MAIS_X
    return d

df_inst['distancia_plot'] =
df_inst['distancia_media'].apply(agrupar_distancia)
ticks_x = list(range(0, LIMITE_X + 1, PASSO_X)) + [POSICAO_MAIS_X]
dominio_x = [-150, POSICAO_MAIS_X + 150]
label_expr_x = f"datum.value == {POSICAO_MAIS_X} ? '2.500+' :
format(datum.value, ',.0d')"

LIMITE_Y = 200
PASSO_Y = 10
POSICAO_MAIS_Y = LIMITE_Y + PASSO_Y

def agrupar_quantidade(q):
    if q > LIMITE_Y:
        return POSICAO_MAIS_Y
```

```
    return q

df_inst['qtd_publicacoes_plot'] =
df_inst['qtd_publicacoes'].apply(agrupar_quantidade)

ticks_y = list(range(0, LIMITE_Y + 1, PASSO_Y)) + [POSICAO_MAIS_Y]
dominio_y = [-10, POSICAO_MAIS_Y + 20]
label_expr_y = f"datum.value == {POSICAO_MAIS_Y} ? '200+' :
format(datum.value, ',.0d')"

pt_br_locale = {"decimal": ",", "thousands": ".", "grouping": [3],
"currency": ["R$", ""]}

chart = alt.Chart(df_inst).mark_circle(
    opacity=0.7,
    stroke='black',
    strokeWidth=1,
    size=400
).encode(
    tooltip=[
        alt.Tooltip('nome_instituicao_destino', title='Instituição:'),
        alt.Tooltip('cidade_instituicao_destino', title='Cidade:'),
        alt.Tooltip('distancia_media', title='Distância (km):',
format=",.0d"),
        alt.Tooltip('qtd_publicacoes', title='Qtde de publicações:',
format=",.0d"),
        alt.Tooltip('pct_endogamia', title='Nível de endogamia:',
format='.0%')
    ],
    x=alt.X('distancia_plot:Q',
            title='Distância (km) da UFSCar - Campus São Carlos',
            scale=alt.Scale(domain=dominio_x),
            axis=alt.Axis(
                values=ticks_x,
                labelExpr=label_expr_x
            )
    ),
    y=alt.Y('qtd_publicacoes_plot:Q',
            title='Quantidade de publicações',
            scale=alt.Scale(domain=dominio_y),
            axis=alt.Axis(
                values=ticks_y,
                labelExpr=label_expr_y
            )
    ),
    color=alt.Color('pct_endogamia:Q',
                    title='Endogamia',
                    scale=alt.Scale(scheme='reds', domain=[0, 1]),
                    legend=alt.Legend(format=".0%")
    )
).properties(
    title='Total de publicações endogâmicas por Instituição',
    width=1000,
    height=1200
).configure(
    locale={"number": pt_br_locale}
).interactive()

chart.save(saida_html)

print(f"\nVisualização salva com sucesso em: {saida_html}")
```

APÊNDICE S – *Script* para visualização do total de publicações não endogâmicas das instituições

Este *script* possibilita a análise da relação entre a distância e a quantidade de publicações não endogâmicas das instituições. A distância tem como referência de origem a UFSCar – Campus São Carlos.

```
import pandas as pd
import altair as alt

arquivo_csv = "coordenadas.csv"
saida_html = "instituicao_distancia_sem_endogamia.html"

csv_params = {'sep': ";", 'dtype': {'codigo_instituicao_destino': str},
'on_bad_lines': 'warn'}

try:
    df = pd.read_csv(arquivo_csv, encoding="utf-8", **csv_params)
except:
    df = pd.read_csv(arquivo_csv, encoding="latin1", engine='python',
on_bad_lines='skip', sep=";", dtype={'codigo_instituicao_destino': str})

df = df.dropna(subset=['codigo_instituicao_destino'])

df["endogamica"] = df["endogamica"].astype(str).str.strip().str.upper()

ids_com_alguma_endogamia = df.loc[df['endogamica'] == 'S',
'id_producao'].unique()
df = df[~df['id_producao'].isin(ids_com_alguma_endogamia)].copy()

df["is_endogamica"] = (df["endogamica"] == 'S').astype(int)

df_inst = df.groupby(['codigo_instituicao_destino',
'nome_instituicao_destino', 'cidade_instituicao_destino']).agg(
    distancia_media=('distancia_km', 'mean'),
    qtd_publicacoes=('id_producao', 'nunique'),
    total_registros=('id_producao', 'count'),
    qtd_endogamicas=('is_endogamica', 'sum')
).reset_index()

df_inst['pct_endogamia'] = df_inst['qtd_endogamicas'] /
df_inst['total_registros']

LIMITE_X = 2500
PASSO_X = 250
POSICAO_MAIS_X = LIMITE_X + PASSO_X
df_inst['distancia_plot'] = df_inst['distancia_media'].apply(lambda d:
POSICAO_MAIS_X if d > LIMITE_X else d)
ticks_x = list(range(0, LIMITE_X + 1, PASSO_X)) + [POSICAO_MAIS_X]
label_expr_x = f"datum.value == {POSICAO_MAIS_X} ? '2.500+' :
format(datum.value, ',.0d')"

LIMITE_Y = 500
PASSO_Y = 25
POSICAO_MAIS_Y = LIMITE_Y + PASSO_Y
df_inst['qtd_publicacoes_plot'] = df_inst['qtd_publicacoes'].apply(lambda
q: POSICAO_MAIS_Y if q > LIMITE_Y else q)
ticks_y = list(range(0, LIMITE_Y + 1, PASSO_Y)) + [POSICAO_MAIS_Y]
```

```
label_expr_y = f"datum.value == {POSICAO_MAIS_Y} ? '500+' :
format(datum.value, ',.0d')"

pt_br_locale = {"decimal": ",", "thousands": ".", "grouping": [3],
"currency": ["R$", ""]}

chart = alt.Chart(df_inst).mark_circle(
    opacity=0.7,
    stroke='black',
    strokeWidth=1,
    size=400
).encode(
    tooltip=[
        alt.Tooltip('nome_instituicao_destino', title='Instituição:'),
        alt.Tooltip('cidade_instituicao_destino', title='Cidade:'),
        alt.Tooltip('distancia_media', title='Distância (km):',
format=",.0d"),
        alt.Tooltip('qtd_publicacoes', title='Qtde de publicações:',
format=",.0d")
    ],

    x=alt.X('distancia_plot:Q',
            title='Distância (km) da UFSCar – Campus São Carlos',
            scale=alt.Scale(domain=[-150, POSICAO_MAIS_X + 150]),
            axis=alt.Axis(values=ticks_x, labelExpr=label_expr_x)
    ),

    y=alt.Y('qtd_publicacoes_plot:Q',
            title='Quantidade de publicações',
            scale=alt.Scale(domain=[-25, POSICAO_MAIS_Y + 50]),
            axis=alt.Axis(values=ticks_y, labelExpr=label_expr_y)
    ),

    color=alt.Color('pct_endogamia:Q',
                    scale=alt.Scale(scheme='reds', domain=[0, 1]),
                    legend=None
    )
).properties(
    title='Total de publicações não endogâmicas por Instituição',
    width=1000,
    height=1200
).configure(
    locale={"number": pt_br_locale}
).interactive()

chart.save(saida_html)
print(f"\nVisualização salva com sucesso em: {saida_html}")
```

APÊNDICE T – *Script* para visualização dinâmica do nível de endogamia e distância geográfica dos pesquisadores

Este *script* tem por objetivo operacionalizar de forma interativa a análise da relação entre endogamia e distância dos pesquisadores, permitindo filtrar pelo ano da publicação. A distância tem como referência de origem a UFSCar – Campus São Carlos.

```
import pandas as pd
import altair as alt

pt_br_locale = {
    "decimal": ",",
    "thousands": ".",
    "grouping": [3],
    "currency": ["R$", ""]
}

arquivo_csv = "coordenadas.csv"

print("Lendo arquivo CSV...")
try:
    df = pd.read_csv(arquivo_csv, sep=";", encoding="utf-8",
                     dtype={'id_producao': str, 'autor': str, 'coautor':
str})
except FileNotFoundError:
    print(f"Erro: O arquivo '{arquivo_csv}' não foi encontrado na pasta.")
    exit()
except pd.errors.ParserError as e:
    print(f"Erro ao ler o CSV: {e}")
    exit()

df['id_producao'] = df['id_producao'].str.strip()
df['autor'] = df['autor'].str.strip()
df['coautor'] = df['coautor'].str.strip()
df = df.drop_duplicates(subset=['id_producao', 'autor', 'coautor'])
df["ano"] = pd.to_numeric(df["ano"], errors='coerce')
df = df.dropna(subset=["ano"])
df["ano"] = df["ano"].astype(int)
df["endogamica"] = df["endogamica"].str.strip().str.upper()
df["is_endogamica"] = (df["endogamica"] == 'S').astype(int)

df["distancia_km"] =
pd.to_numeric(df["distancia_km"].astype(str).str.replace(',', '.'),
errors='coerce')

df_analise = (
    df.groupby(["id_producao", "autor"])
      .agg({
          "is_endogamica": "mean",
          "distancia_km": "mean",
          "coautor": "nunique",
          "ano": "max"
      })
      .rename(columns={"coautor": "tamanho"})
      .reset_index()
)

df_analise.columns = [
```

```
    "ID",
    "Autor",
    "Nível de endogamia",
    "Distância média",
    "Tamanho",
    "Ano"
]

df_analise["Tamanho"] = df_analise["Tamanho"].astype(float)

df_analise["Distância média"] = df_analise["Distância média"].apply(
    lambda x: 1650 if x > 1500 else x
)

tamanho_base_area = 120
max_coautores = df_analise["Tamanho"].max()
if pd.isna(max_coautores) or max_coautores == 0:
    max_coautores = 1.0
tamanho_max_area = tamanho_base_area * max_coautores

ano_min = int(df_analise["Ano"].min())
ano_max = int(df_analise["Ano"].max())

valores_legenda = [1, 5, 10, 20, 30, 40, 50, 60]
valores_legenda = [v for v in valores_legenda if v <= max_coautores + 5]

print(f"Dados processados. Gerando gráfico para anos {ano_min}-
{ano_max}...")

param_todos = alt.param(
    name="Todos",
    value=True,
    bind=alt.binding_checkbox(name="Todos os anos:")
)

param_ano = alt.param(
    name="Ano",
    value=ano_min,
    bind=alt.binding_range(
        min=ano_min,
        max=ano_max,
        step=1,
        name="Ano:"
    )
)

faixa_1500 = alt.Chart(
    pd.DataFrame({"y1": [1500], "y2": [1800]})
).mark_rect(
    fill="#f5f5f5", stroke="#999999", strokeDash=[5,5], opacity=0.4
).encode(y="y1:Q", y2="y2:Q")

base = (
    alt.Chart(df_analise)
    .mark_circle(
        opacity=0.8,
        stroke="black",
        strokeWidth=0.5
    )
    .encode(
        x=alt.X(
            "Nível de endogamia",
```

```
            title="Nível de endogamia",
            scale=alt.Scale(domain=[-0.05, 1.05], nice=False),
            axis=alt.Axis(
                format="%",
                grid=True,
                values=[0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9,
1.0]
            )
        ),
        y=alt.Y(
            "Distância média",
            title="Distância média (km) da UFSCar – Campus São Carlos",
            scale=alt.Scale(domain=[0, 1800], nice=False),
            axis=alt.Axis(values=[0, 250, 500, 750, 1000, 1250, 1500],
format=",.0f")
        ),
        size=alt.Size(
            "Tamanho:Q",
            title="Qtde de coautores",
            scale=alt.Scale(domain=[1, max_coautores],
range=[tamanho_base_area, tamanho_max_area]),
            legend=None
        ),
        color=alt.Color(
            "Tamanho:Q",
            title="Qtde de coautores",
            scale=alt.Scale(domain=[1, max_coautores], scheme="greenblue",
reverse=True),
            legend=alt.Legend(
                type="gradient",
                gradientLength=200,
                values=valores_legenda
            )
        ),
        tooltip=[
            alt.Tooltip("ID", title="ID da publicação:"),
            alt.Tooltip("Autor", title="ID do autor:"),
            alt.Tooltip("Distância média", title="Distância média (km):",
format=',.0f'),
            alt.Tooltip("Tamanho", title="Qtde de coautores:",
format=".0f"),
            alt.Tooltip("Nível de endogamia", title="Nível de endogamia:",
format='.0%')
        ]
    )
    .add_params(param_todos, param_ano)
    .transform_filter((param_todos) | (alt.datum.Ano == param_ano))
)

chart = (faixa_1500 + base).properties(
    title=f"Endogamia por Pesquisador ({ano_min}-{ano_max})",
    width=800,
    height=600
).configure_axis(
    gridColor="#EAEAEA", titleFontSize=14, labelFontSize=12
).configure(
    locale={"number": pt_br_locale}
)

nome_arquivo_saida = "pesquisador_endogamia_distancia.html"
chart.save(nome_arquivo_saida)
print(f"\nSucesso! Visualização salva como: {nome_arquivo_saida}")
```