23° enancib
ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO
06 a 10 de novembro de 2023
ARACAJU-SE

GT 7 – Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação

ISSN 2177-3688

O USO DE TÉCNICAS DE CIÊNCIA DE DADOS PARA ANALISAR A AMBIGUIDADE DE AUTORIA EM PRODUÇÃO CIENTÍFICA DOS PROFESSORES DOS PROGRAMAS DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO DAS UNIVERSIDADES FEDERAIS BRASILEIRAS

THE USE OF DATA SCIENCE TECHNIQUES TO ANALYZE THE AMBIGUITY OF AUTHORSHIP IN SCIENTIFIC PRODUCTION OF TEACHERS OF POSTGRADUATE PROGRAMS IN INFORMATION SCIENCE OF BRAZILIAN FEDERAL UNIVERSITIES

Jéssica Bilac Gaspareto – Universidade de Brasília (UNB)
Marcio de Carvalho Victorino – Universidade de Brasília (UNB)

Modalidade: Trabalho Completo

Resumo: O objetivo deste trabalho é analisar a ambiguidades entre os nomes dos professores dos programas de pós-graduação em Ciência da Informação das universidades federais brasileiras. Desta forma, tem como objetivos específicos a contextualização do ciclo da comunicação cientifica, seus pilares e infraestruturas; como a colaboração científica funciona; importância da autoria múltipla; além de apresentar como as técnicas de ciência de dados podem atuar no contexto de ambiguidade em produção científica. A metodologia adotada é quantitativa, partindo da coleta de dados proveniente do Portal Sucupira da CAPES, plataforma Lattes e plataforma ORCID. Em seguida, os dados coletados foram armazenados em um esquema de banco de dados relacional criado no sistema gerenciador de banco de dados MySQL, que possibilitou a organização, manipulação e limpeza dos dados. Após essas etapas, esses dados foram utilizados para o desenvolvimento do estudo de caso apresentado neste trabalho e por meio da plataforma Tableau (plataforma utilizada para análise de dados), foram gerados gráficos para discussão do problema apresentado. Foi possível verificar dois tipos diferentes de ambiguidades entre o conjunto de autores formado por professores participantes dos programas de pós-graduação em Ciência da Informação nas universidades federais brasileiras, sendo elas *mixed citation* e *split citation*. Os dois tipos de ambiguidades encontrados no banco de dados acabam acarretando problemas de mensurações em métricas para indicadores de qualidade, afetando os índices de produções científicas, sendo consideradas objetos de estudos significantes para pesquisas dentro da Ciência da Informação. A análise inicial desta pesquisa, que envolvia a identificação da ambiguidade entre os professores dos programas de pós-graduação em Ciência da Informação das universidades federais brasileiras, já foi finalizada, enquanto a segunda fase está atualmente em andamento, com planos de publicação no futuro.

Palavras-chave: comunicação Científica; ciência de dados; ambiguidade entre autoridades; colaboração científica; comunidade científica.

Abstract: The objective of this study is to analyze the ambiguities among the names of professors in graduate programs in Information Science at Brazilian federal universities. Thus, it has specific objectives such as contextualizing the cycle of scientific communication, its pillars, and infrastructures; how scientific collaboration works; the importance of multiple authorship; as well as demonstrating how data science techniques can operate in the context of ambiguity in scientific production. The methodology adopted is quantitative, starting with data collection from the CAPES Sucupira Portal, Lattes platform, and ORCID platform. Next, the collected data were stored in a relational database

schema created in the MySQL database management system, enabling data organization, manipulation, and cleaning. After these stages, this data was used for the development of the case study presented in this work, and through the Tableau platform (a platform used for data analysis), graphs were generated to discuss the problem at hand. Two different types of ambiguities were identified among the group of authors composed of professors participating in graduate programs in Information Science at Brazilian federal universities, namely mixed citation and split citation. Both types of ambiguities found in the database end up causing problems in measuring metrics for quality indicators, affecting scientific production indices, and are considered significant objects of study for research within the field of Information Science. The initial analysis of this research, which involved identifying ambiguity among professors in graduate programs in Information Science at Brazilian federal universities, has already been completed, while the second phase is currently ongoing, with plans for future publication.

Keywords: scientific communication; data science; ambiguity between authorities; scientific collaboration; scientific community.

1 INTRODUÇÃO

A ascendência de novas descobertas e avanços científicos foram fundamentais para transformações, em todas as épocas, por meio de mudanças de padrões de comportamento e do acesso à informação na sociedade. Juntamente com o advento da Revolução Industrial e a chegada do século XX – marcado pelo impulso sem precedentes do conhecimento e desenvolvimento tecnológico –, a ciência ganhou maior protagonismo juntamente com a produção de novas informações, fazendo com que diversos pesquisadores necessitassem publicar seus trabalhos, nascendo assim a comunicação científica (VALEIRO; PINHEIRO, 2008).

A comunicação científica envolve todo um processo de produção e disseminação da informação que, necessita de uma forma de registro para conceder os direitos autoriais a um indivíduo, uma vez que, um pesquisador dedicou esforço e tempo para contribuir para o avanço da ciência.

Uma das formas para mensurar a qualidade de uma produção científica é feito pela atribuição de indicadores de qualidade, ressaltando como um dos principais a autoria múltipla. A autoria múltipla é um indicador que exibe como a contribuição de autores em projetos científicos geram mais reconhecimento e tende a proporcionar um melhor avanço na ciência.

A ambiguidade entre nome de autores, onde um autor pode ter uma enorme gama de variações em seu nome ou diferentes autores com o mesmo nome, podem ser um fator crítico para a análise do indicador de qualidade em autoria múltipla por conta de duplicatas que acarretam na imprecisão dos resultados.

Neste contexto, por meio de produções científicas e a atribuições de autoridades em trabalhos científicos, é possível identificar padrões de citações para exibir a problemática deste trabalho uma vez que ao analisar as produções científicas, sejam elas produtos de colaboração científica ou produções individuais, nota-se ambiguidades entre autoridades dificultando na identificação dos autores para reconhecimento dos créditos acerca de uma produção científica.

A pergunta principal desta pesquisa baseia-se em como analisar a ambiguidade de autoria considerando a produção científica dos professores de programas de pós-graduação em Ciência da Informação das universidades federais brasileiras.

O objetivo deste trabalho é elaborar e sistematizar uma análise envolvendo uma amostra dos professores dos programas de pós-graduação em Ciência da Informação das universidades federais brasileiras para exibir casos de ambiguidade em autoridades desses professores. Este estudo apresentará por meio de gráficos e a exibição de um modelo de dados relacional alguns casos de ambiguidades encontrados na base construída.

Dessa forma, as técnicas utilizadas em ciência de dados podem ser um fator chave para tratar e limpar os dados que envolvem a autoria múltipla, reduzindo ou eliminando as duplicatas e trazendo precisão para os resultados de busca de autores. As técnicas de ciência de dados também podem ser utilizadas para exibir gráficos e números estatísticos para representar o problema apontado.

2 DESENVOLVIMENTO

A comunicação dispõe de um papel central na ciência pelo fato de que, para ser considerado científico, um determinado conhecimento necessita da aprovação de outros pesquisadores. Essa aprovação se dá em dois momentos – o primeiro ocorre antes da publicação por meio de um teste de qualidade denominado “avaliação prévia[1]” e o segundo ocorre após a publicação, quando é aprovado na avaliação prévia, sendo publicado como artigo científico e exposto à crítica de todos –, ao ser publicado e acessível aos demais pesquisadores, esse conhecimento pode contribuir para outras pesquisas gerando novos conhecimentos (MUELLER, 2007).

[1] É o processo de julgamento que um manuscrito é submetido antes de uma publicação realizada pelos pares (MUELLER, 2007).

Uma vez que esses conhecimentos são publicados, a comunicação científica nasce mediante a inevitabilidade do registro dos quais os avanços científicos e tecnológicos produzidos pelo ser humano sucedem, isto ocorre para que os precursores de diversas áreas do conhecimento possam prosperar daquela bagagem científica, contribuindo para pesquisas pelo bem da humanidade.

Segundo Targino (2000), a comunicação científica torna-se indispensável no âmbito das atividades científicas, pois possibilita a conexão de esforços individuais de membros da comunidade científica, e então, viabilizando uma troca contínua de informações e difundindo conhecimentos para sucessores ou auferidos de seus predecessores.

O processo da comunicação científica pode ser considerado como um sistema cíclico, pois precisa passar por algumas etapas que devem se repetir para a garantia do avanço científico e retroalimentação deste ciclo.

2.1 Ciência da Informação e Comunicação Científica

Conforme Miranda (2002), a Ciência da Informação teve seu advento após o fenômeno da "explosão da informação", situada após a 2°Guerra Mundial, e na necessidade de um controle bibliográfico resultando no tratamento da documentação implícita no processo.

O termo conhecido como "Ciência da Informação" ou "CI", foi criado por volta de 1960, com base em estudos de produção, processamento e uso da informação como atividade predominantemente humana. Todavia, Wellish (1987) assegura que o termo "Ciência da Informação" foi utilizado pela primeira vez em 1959, para caracterizar o estudo do conhecimento registrado (HELPRIN, 1989, *apud* PINHEIRO; LOUREIRO, 1995).

Ao longo de 1962, um grupo de pesquisadores reunidos no *Georgia Institute of Technology* declararam que a Ciência da Informação era a ciência que investigava o comportamento da informação, suas propriedades, forças que regem o fluxo da informação e por fim seus meios de processamento visando o melhor uso da informação. É uma área que se relaciona e deriva de outras áreas, por isso é considerada uma área emergente de novas disciplinas interdisciplinares (ROBREDO, 2003; VICTORINO 2011).

No decorrer da década de 1990, a Ciência da Informação é definida como um campo dedicado às questões científicas com foco em práticas profissionais voltadas aos problemas em relação à comunicação do conhecimento registrado entre os seres humanos (SARACEVIC, 1992).

A Ciência da Informação por se tratar de uma área do conhecimento com foco nas questões científicas se alinha diretamente com a comunicação científica uma vez que, ao falar dos centros de interesses e de ações da Ciência da Informação, destaca-se ao que se refere a comunicação científica – conhecer os tipos de publicações, características e formas –, mas havendo a necessidade de compreender também as características próprias da informação científica – sua estrutura de processos e seus sistemas de comunicação –, alinhando ambas em prol do avanço científico (MUELLER, 2007).

A comunicação científica é considerada parte essencial referente aos estudos da Ciência da Informação, na qual compõe uma disciplina cujo encargo central é atribuído a questões relacionadas – direta ou indiretamente – com o compartilhamento do conhecimento na sociedade (BAPTISTA *et al.*, 2007).

2.2 Colaboração Científica

De acordo com Grácio (2018), a coautoria no ramo científico, ainda que em pequenas quantidades, já ocorria no século XVII, tendo o primeiro registro de artigo escrito contando com coautoria entre pesquisadores na data de 1665.

Dessa forma, pensando o produto da ciência como um fator de avanço à sociedade, a colaboração científica é definida por um esforço cooperativo que busca alcançar metas em comum entre os pesquisadores, esforço coordenado e resultados – os trabalhos científicos –, por meio de méritos e responsabilidades compartilhadas (BALANCIERI *et al.*, 2005). Nesta linha de pensamento, cientistas que trabalham em conjunto podem acelerar suas pesquisas por trabalharem em alinhamento com outros pesquisadores ao buscarem o mesmo, resultando no enriquecimento da ciência.

A colaboração científica acaba potencializando o crescimento profissional de um pesquisador, uma vez que, permite o trabalho conjunto de pesquisadores mais renomados com pesquisadores iniciantes (GRÁCIO, 2018). Este crescimento profissional gerado pode ser estipulado com base em novas produções científicas e publicações em revistas que assim, eventualmente, servirão de base para contribuição de novos precursores a fim de renovar e manter o conhecimento científico.

Segundo Katz e Martin (1997), a colaboração científica ocorre com diferentes públicos – nações, instituições, grupos de pesquisas –, assim potencializando os resultados quando um grupo emprega esforços para chegar no mesmo objetivo. Um dos grandes exemplos para

colaboração científica em nações pôde ser apresentado durante a pandemia do Covid-19, iniciada no Brasil em 2020, onde cientistas de todo o mundo se mobilizaram para estudar formas de combate à pandemia ressaltando o potencial da colaboração científica.

2.3 Ambiguidade de Autoria em Produções Científicas

Os programas de pós-graduação brasileiros inseridos na CAPES necessitam periodicamente que seja feito um levantamento das produções acadêmicas dos pesquisadores, grupos de pesquisa, projetos, entre outros presentes na plataforma de avaliação da CAPES (BRAUNER; ARAÚJO; SANTOS, 2016).

De acordo com Brauner, Araújo e Santos (2016), durante esses levantamentos, recomenda-se que haja um olhar mais crítico acerca das ambiguidades presentes ao alinhamento realizado por softwares. Essas ambiguidades costumam afetar publicações, nomes de eventos e o mais relevante para esta pesquisa, a ambiguidade entre nome de autores.

Conforme Mugnaini *et al* (2012), essa problemática acerca da ambiguidade entre nome de autores surge devido ao fato de que, um sistema de Ciência e Tecnologia de qualquer instituição ou país passa por uma avaliação de sua produção científica. Essa produção necessita estar devidamente inserida e indexada em uma base de dados normalizada, diversificando olhares e indicadores, favorecendo sua apresentação e classificação.

Lee *et al.* (2005) afirmam que o problema da ambiguidade de nomes pode haver mais um desdobramento, sendo classificados em dois subproblemas: *split citation* – quando um autor possui diversas variações em seu nome e *mixed citation* – quando autores diferentes possuem nomes iguais –, dessa forma agravando mais a complexidade da desambiguação.

A ocorrência do *split citation* pode ser demonstrada na Figura 1, onde um mesmo autor possui a variação de oito formas para ser citado. Esse tipo de problema é comum ocorrer com autores com sobrenomes incomuns ou quando possuem nomes extensos. É importante ressaltar que o exemplo apresentado na Figura 1 trata-se de um exemplo de um caso real dentro da Plataforma Sucupira.

Figura 1 - Exemplo de *Split Citation*

Abreviaturas:	Simeao, E. SIMEAO, E. SIMAO, E. L. M. S. SIMEAO, E. L. M. S. SIMEÃO, E. L. M. S. **(Principal)** SIMEAO, E. L. SIMEAO, ELMIRA LUZIA MELO SOARES SIMEAO, Elmira L. M. S.

Fonte: Elaborado pelos autores (2021).

Sobre a outra ocorrência apresentada por Lee *et al*. (2005), *mixed citation*, é apresentada na Figura 2. A forma de ambiguidade *mixed citation* pode ser ocasionada pela forma como uma revista acaba indexando e citando o nome de um autor. O sobrenome Souza é exibido em diversos resultados da busca por autores com esse nome devido ao fato de ser um sobrenome comum. Nesse tipo de ocorrência é bastante comum a ambiguidade acerca da autoria visto que, ao utilizar apenas o sobrenome Souza e um nome, haverá diversos outros autores com o mesmo nome causando imprecisão na desambiguação. Em síntese, quando dois autores tiverem o mesmo nome e sobrenome, como consequência, o nome de autoridade apresentado será o mesmo, necessitando de novos componentes para diferenciação e desambiguação de dados autorais, como por exemplo o número de ORCID de cada autor, qual instituição aquele professor é filiado e outros elementos.

Figura 2 - Exemplo de *Mixed Citation*

- João Carlos Félix Souza
- João E. Tozzi de Souza
- João Henrique inácio de Souza
- João MO de Souza
- João Marcelo Souza
- João N. de Souza
- João Nunes Souza
- João Nunes de Souza

Fonte: Elaborado pelos autores (2021).

De acordo com Mugnaini *et al.* (2012), as causas da ambiguidade entre nome de autores, geralmente, ocorrem no próprio documento quando o autor se autodenomina de variadas formas em diferentes momentos (nome por extenso, abreviado, nome de casado,

fantasia ou outras variações) ou quando as bases de dados geram de forma automática a forma nominal para citação.

A ambiguidade da autoridade torna-se um problema para análises bibliométricas, duplicidades em repositórios digitais e um dos mais críticos, na garantia de créditos em métricas de citações para pessoas errôneas.

2.4 Ciência de Dados

A Ciência de Dados *(data science)*, pode ser considerada como um reflexo do ambiente interconectado com sua enorme quantidade de dados disponíveis aos quais conhecemos. Trata-se de uma área interdisciplinar que pode abranger áreas como as: ciências exatas e engenharias (COMARELA *et al*., 2019).

A ciência de dados teve sua ascensão com o advento do desenvolvimento das tecnologias de informação e das possibilidades de busca, por meio de mecanismos mais aprimorados como buscas avançadas (REIS, 2019).

Reis (2019) afirma que, a ciência de dados é considerada a ciência que reúne múltiplos aspectos de informações por meio de dados e contando com uma equipe multidisciplinar envolvendo profissionais de diversas áreas como: matemáticos, estatísticos, programadores, analistas de dados e bibliotecários.

Apesar do termo "ciência de dados" ser relativamente novo, a busca para compreensão desses dados por meio do trabalho de estatísticos, cientistas e profissionais da informação, já vinham sendo abordadas em um espaço de discussões antigas (PRESS, 2013; ROLIM, 2018).

Analisando uma evolução histórica elaborada por Press (2013), é possível perceber como a ciência de dados teve sua evolução ao longo das últimas décadas. No Quadro 1 é proposto um modelo fundamentado nessa cronologia apresentando os princípios norteadores da atual ciência de dados:

Quadro 1 - Modelo Cronológico dos Princípios Norteadores da Atual Ciência de Dados

Década de 1960	Surgimento de livros, publicações seriadas, artigos, *workshops*, entidades e encontros de especialistas que abordam análise de dados e *big data;*
Em 1966	Lançamento do livro *from data mining to knowledge Discovery in databases* de Usama Fayyad, Gregory Piatetsky-Shapiro e Padhraic Smyth; o periódico *The Journal Data Mining and Knowledge Discovery*, lançado em 1997;
Em 1977	O livro *Exploratory data analysis*, em 1977, de John W. Tukey;
Em 2002	O periódico *Data Science Journal*;
Em 2005	O livro *Competing on analytics* de Thomas H. Davenport, Don Cohen e Al Jacobson;

Em 2009	O artigo *Rise of the data scientist*;
Em 2010	O artigo *What is Data Science?* de Mike Loukides;
Em 2012	O artigo *Data scientist: the sexiest job of the 21st century* de Thomas H. Davenport e D. J. Patil;

Fonte: Elaborado pelos autores (2021), com base em Press (2013) e Rolim (2018, p. 38).

3 METODOLOGIA

A pesquisa configura-se como quantitativa devido ao fato de que, é uma pesquisa que utiliza a estatística para a mensuração do fenômeno de ambiguidade entre os professores presentes nos programas de pós-graduação das universidades federais brasileiras em Ciência da Informação (SAMPIERI *et al*., 2013). Por meio dos métodos comprobatórios envolvendo este estudo, é possível demonstrar através dos resultados dois tipos de casos de ambiguidade entre os professores.

Sobre o procedimento adotado para este trabalho, utilizou-se a pesquisa bibliográfica para entendimento de conceitos chaves para este estudo e conhecimento do estado da arte do tema escolhido. Este levantamento bibliográfico foi realizado por meio de: livros, bases de dados de acesso restrito e livre. Desta forma, a parte bibliográfica foi essencial para construção do modelo conceitual.

Os buscadores utilizados na pesquisa foram: “autoria múltipla”, “coautoria”, “métodos para desambiguação de autores”, “bibliometria para desambiguação”, “uso da ciência de dados para desambiguação”, “autoridade de autores”, “comunicação científica”, “autoria múltipla como indicador de qualidade científica”.

Após entendimento e construção da parte teórica deste trabalho por meio do levantamento bibliográfico, foi pensando na parte da pesquisa aplicada que, inicialmente, foi criado em modelo conceitual para representação do domínio da autoria em produção científica dos professores dos programas de pós-graduação em Ciência da Informação das universidades federais brasileiras. Esse modelo conceitual foi mapeado para um esquema lógico relacional de dados com as respectivas tabelas que foram criadas no sistema gerenciador de banco de dados MySQL para a persistências dos dados. Posteriormente, foi pensado quais bases de dados possuíam, os dados necessários ao estudo.

Foram extraídos dados da Plataforma Sucupira no Portal da Capes de forma manual sobre as universidades que possuem programas de pós-graduação da área de Ciência da Informação no Brasil e respectivos professores. Sugere-se para futuras pesquisas o uso de *scripts* em Python para realizar o web *scrapping* de forma automática.

Em seguida, esse levantamento na Plataforma Sucupira, foram acessados os perfis dos professores na plataforma Lattes com a finalidade de extrair os números de ORCID, sendo importante frisar que nem todos os professores tinham o ORCID vinculado no Lattes, neste caso, criou-se um número aleatório para considerar como chave única na ausência do ORCID. Para levantamento das variações de nomes dos professores (autores), foi utilizada a própria plataforma Sucupira que disponibiliza todas as variações disponíveis utilizadas pelos autores.

Após a extração e organização de todos esses dados apresentados, foi desenvolvido uma base de dados relacional que deu origem ao estudo de caso composto por um esquema relacional criado no sistema gerenciador de banco de dados (SGBD) MySQL.

O esquema relacional criado tem por objetivo oferecer a organização e ordenação dos dados, para a criação de consultas utilizando a linguagem SQL e visualização dos dados utilizando a ferramenta Tableau[2], a fim de apresentar por meio de gráficos analíticos, a mensuração da ambiguidade entre autores.

4 ESTUDO DE CASO

Para este estudo de caso foram utilizados os dados provenientes da Plataforma Sucupira do Portal da CAPES, em razão da plataforma conter dados atualizados e correntes acerca dos programas de pós-graduações presentes em universidades federais do Brasil e seus respectivos docentes participantes.

O presente trabalho tem como objetivo exibir a ambiguidade em nomes de autores vigentes dos programas de pós-graduações presentes na base de dados da Plataforma Sucupira no ano de 2020. Para o desenvolvimento deste estudo de caso foram seguidos os seguintes passos:

- Criação de um modelo conceitual de dados para representar o domínio observado;
- Levantamento de fontes de dados que possuem os dados representados no modelo de dados;
- Mapeamento do modelo conceitual para o modelo lógico relacional com a definição das respectivas tabelas;

[2] Trata-se de uma ferramenta paga, porém, possuí plano gratuito para e-mails acadêmicos ou estudantis. https://www.tableau.com/pt-br. Acesso em 11 set. 2023.

- Criação de um esquema no sistema gerenciador de banco de dados MySQL com as tabelas modeladas;
- Extração dos dados para a carga das tabelas criadas;
- Apresentação desses dados por meio de gráficos analíticos utilizando a ferramenta Tableau.

4.1 Criação do Modelo Conceitual

Nesta primeira etapa foi proposto um modelo conceitual baseado na técnica proposta por Peter Chen (1976).

Para a criação do modelo conceitual, foi pensado em um modelo de dados capaz de representar os seguintes requisitos:

A. O banco de dados deve comportar professores (autores) de cada programa de pós-graduação presentes na Plataforma Sucupira;

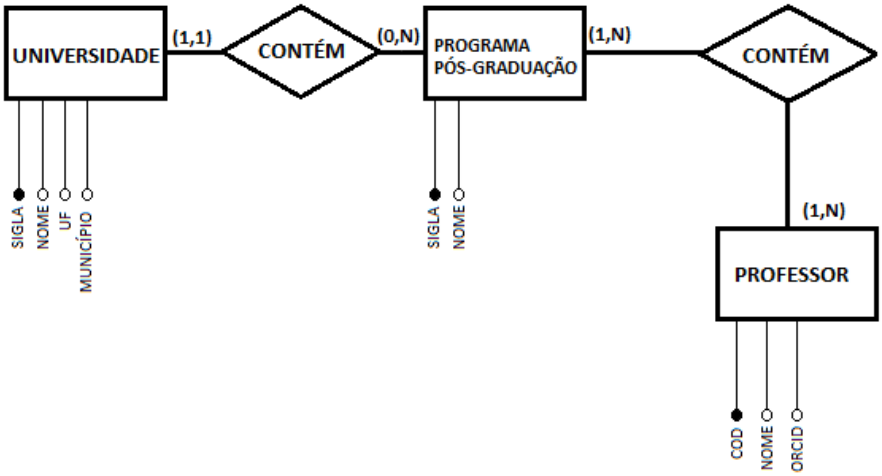
B. O banco deve comportar os programas de pós-graduações presentes em cada universidade federal (sendo que algumas universidades possuem mais de um programa de pós-graduação em sua instituição);

C. O banco deve comportar cada universidade que conta com área correlatas acerca de Ciência da Informação;

D. O banco deve comportar uma entidade acerca das variações de nomes de cada professor (autor);

O modelo conceitual está apresentado na Figura 3, nesse protótipo ainda não havia sido pensado em uma entidade para inserção dos dados acerca das variações de nomes dos professores. A Figura 4 apresenta esse modelo conceitual mapeado para o nível lógico utilizando a ferramenta *workbench* do sistema gerenciador de banco de dados MySQL.

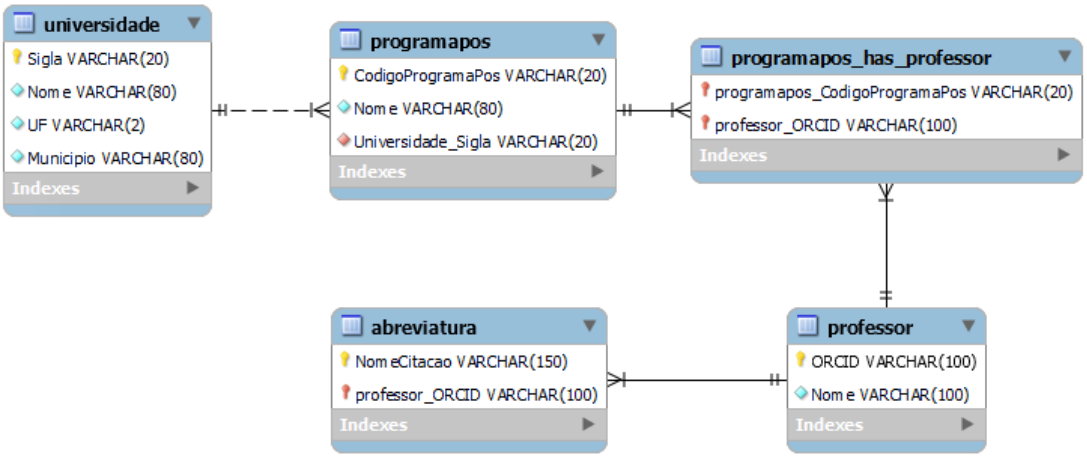
Figura 3 - Modelo Conceitual Inicial



Fonte: Elaborado pelos autores (2021).

Após a modelagem apresentada, foi pensado em mais uma entidade para representar as variações dos nomes dos autores. Na Figura 4 é apresentado o modelo lógico do banco de dados:

Figura 4 - Modelo Final Lógico



Fonte: Elaborado pelos autores (2021).

4.2 Levantamento de fontes de dados que possuem os dados representados no modelo de dados

Nesta parte do estudo de caso foram selecionadas as fontes de dados que possuem os dados representados no modelo lógico apresentado na seção anterior, levando em consideração os requisitos apresentados.

4.3 Fonte dos Dados para Professores

Para a entidade professores (autores), os dados vieram da Plataforma Sucupira do Portal da CAPES. Os requisitos para coleta de professores foram os seguintes:

- Ano: 2020
- Apenas as instituições de ensino superior presentes na Plataforma Sucupira e presentes em programas de pós-graduações acerca de Ciência da Informação;
- Categoria de condições de professores permanentes e colaboradores;

Para a criação de uma chave primária, foi necessário pensar em um campo que seria único para cada docente, ou seja, este número não poderia variar e nem se repetir. Posteriormente, foi utilizado o ORCID de cada professor como sua chave primária. Para os professores que não possuem ORCID utilizou-se a contagem numérica como identificador.

4.4 Fonte dos Dados para Universidades e Programas de Pós-Graduação

Para a entidade universidade, os dados vieram da Plataforma Sucupira do Portal da CAPES. Os requisitos para coleta de dados foram os seguintes:

- Cursos avaliados e conhecidos da área de Comunicação e Informação;
- Dentro da área de Comunicação e Informação foram coletadas apenas os cursos da área de Ciência da Informação, excluindo os outros cursos (comunicação, desenho industrial, e museologia);

Foram identificados pelo código disponível de cada programa de pós-graduação para utilizar como sua chave primária;

4.5 Definição dos Atributos Persistentes para fins de Coleta de Dados

Durante esta etapa foram analisados quais dados seriam relevantes para extrações futuras e quais seriam persistidos no sistema gerenciador de banco de dados MySQL (SGBD).

Para cada entidade foram definidos os seguintes atributos:

- **Entidade Professor**: ORCID, Nome; Universidade

- **Entidade Programa Pós-Graduação**: Código Programa Pós, Nome e Sigla da Universidade;
- **Entidade Universidade**: Sigla, Nome, UF e Município;
- **Entidade Abreviatura**: Nome para Citação e ORCID;

Para os relacionamentos com chaves estrangeiras foram definidos os seguintes atributos:

- **Relacionamento Programa Pós-Graduação e Professor**: Código Programa Pós e ORCID Professor;

Na Tabela 1 apresentada abaixo é sintetizado cada uma das entidades e relacionamentos com seus totais de registros:

Tabela 1 - Quantidade de Registros em Cada Tabela

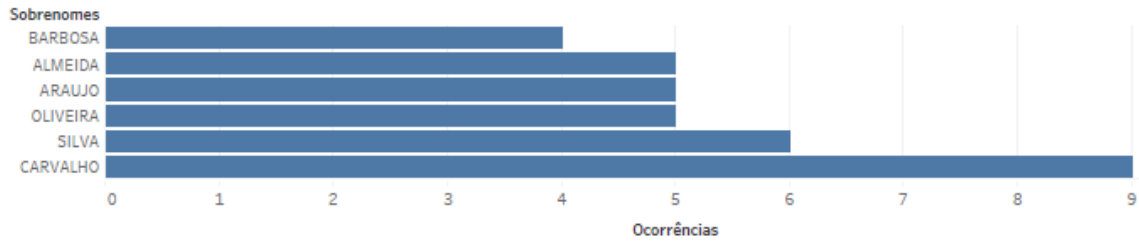
Professor	443 registros
Universidade	23 registros
Programas Pós	26 registros
Abreviatura	1342 registros
Relacionamento Estrangeiro (Professor e Programa Pós)	418 registros

Fonte: Elaborado pelos autores (2021).

5 RESULTADOS

No Gráfico 1, observa-se que o sobrenome com maior número de ocorrências no banco de dados foi ‘Carvalho’ com nove registros de autores diferentes. Este gráfico exclui as duplicatas de sobrenome de um mesmo autor para evidenciar apenas professores diferentes com o mesmo sobrenome.

Gráfico 1 - Sobrenomes com maiores Ocorrências

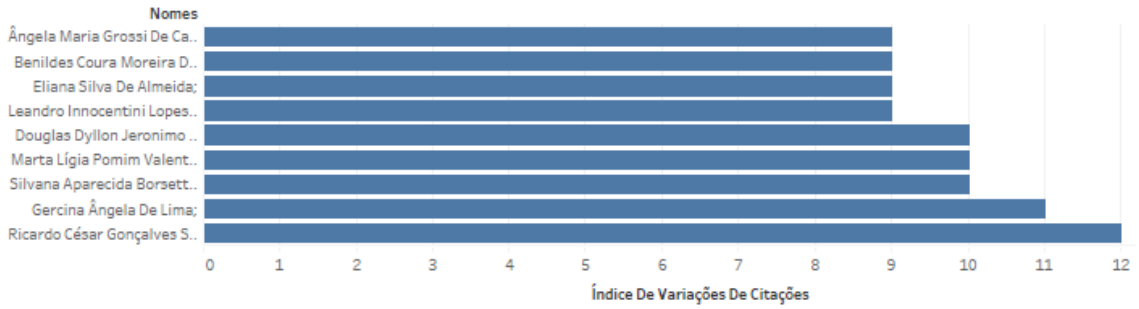


Fonte: Elaborado pelos autores (2021).

Este caso é um exemplo de *mixed citation* apontado por Lee *et al.* (2005), evidenciando que caso um professor opte por utilizar apenas o sobrenome ‘Carvalho’ e ‘ano’ para citações, pode haver ambiguidades com outros professores registrados em bases de dados.

No Gráfico 2, observa-se que este gráfico representa os nove autores com maiores números de variações de nomes para citações. Este é um caso de *split citation* apontado por Lee *et al.* (2005), cujo autor possui diversas formas de ser citado, sendo este problema causado por autores que possuem sobrenomes extensos ou incomuns. O professor Ricardo Sant'Ana foi considerado o exemplo mais pertinente desse tipo de ocorrência contando com 12 formas de abreviatura para citações.

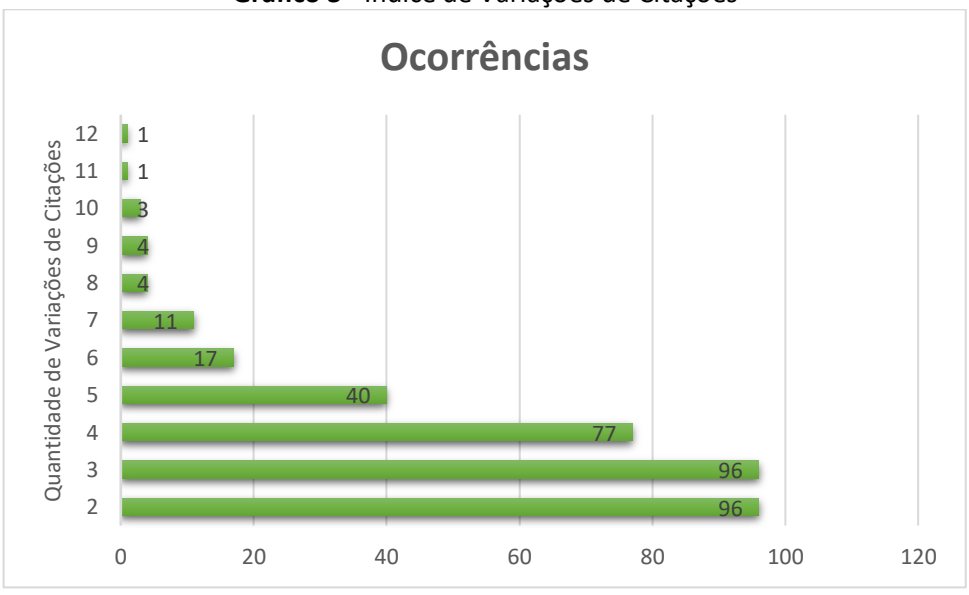
Gráfico 2 - Autores com Maiores Índices de Variações em Citações



Fonte: Elaborado pelos autores (2021).

No Gráfico 3, observa-se a quantidade de registros dentro do banco de dados (Quantidade de Variações de Citações) em comparação com a quantidade de vezes que foram encontrados esses registros (Ocorrências). Desta forma é possível analisar que os professores com intervalos de 2 até 5 formas de variações foram os mais exibidos no banco de dados totalizando 309 registros.

Gráfico 3 - Índice de Variações de Citações



Fonte: Elaborado pelos autores (2021).

Em decorrência das análises realizadas anteriormente, verificou-se neste trabalho várias ocorrências de *mixed citation* predominantemente em sobrenomes considerados comuns nos professores inseridos no banco de dados relacional, sendo os sobrenomes com maiores ocorrências exemplificadas no Gráfico 1.

A ocorrência de *split citation* foi exibida na maioria das variações nominais dos professores presentes no banco de dados relacional. O Gráfico 3 exemplifica esta ocorrência durante os intervalos de variações nominais, onde considera-se um valor acima da média, na ocorrência de 5-12 tipos de variações pelo fato de que a maioria dos autores estavam na margem de 2-4 variações.

As UFs que mais concentraram percentual de ambiguidades foram SP (25%), RJ (11%) e o DF (10%), sendo que, o restante das outras UFs ficaram entre 7%-2% no percentual de ambiguidade. Em sequência, as instituições com maiores ambiguidades foram UNESP-MAR (18,23%) e a UNB (10,42%). Por fim, ressalta-se que nomes entre 2-5 variações são os mais frequentes a possuírem ambiguidades.

6 CONSIDERAÇÕES FINAIS

Este trabalho proporcionou a visualização de dois tipos de ambiguidades que costumam aparecer com frequência em base de dados científicas. Essas duas ambiguidades acabam acarretando problemas de mensurações em métricas para indicadores de qualidade que afetam os índices de produções científicas, sendo consideradas objetos de estudos significativos para pesquisas na área da Ciência da Informação.

Por meio das técnicas de ciência de dados, como por exemplo, a extração, limpeza, organização de dados e apresentação desses dados em gráficos analíticos, foi possível a visualização da ambiguidade de autoria em produção científica dos professores dos programas de pós-graduação em ciência da informação das universidades federais brasileiras presentes na base de dados relacional disponibilizada.

Essa base de dados foi carregada com dados provenientes da Plataforma Sucupira do Portal de Periódicos da CAPES e dados captados de outras plataformas.

As maiores dificuldades encontradas no desenvolvimento deste trabalho foram aquelas relacionadas à coleta dos dados, por se tratar de dados que precisavam passar por uma limpeza, validação e inserção manual no banco de dados relacional.

Outro ponto considerado uma dificuldade para a progressão do trabalho se deu por meio de definir quais atributos seriam pertinentes para cada tabela modelada, para proporcionar as eventuais consultas. Desta forma, após definir quais atributos definitivos integrariam o banco de dados relacional, foram realizadas consultas SQL que possibilitaram carregar as tabelas para proporcionar a análise das ambiguidades por meio de gráficos utilizando a ferramenta Tableau, confirmando a existência do problema previamente apresentado.

Este trabalho apresenta como sugestão para trabalhos futuros a criação de um algoritmo ou *software* capaz de realizar o tratamento e resolução, de forma automatizada, de ambiguidades a partir de dados coletados de bases de dados.

REFERÊNCIAS

BALANCIERI, R.; BOVO, A. B.; KERN, V. M.; PACHECO, R. C. D. S.; BARCIA, R. M. A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na Plataforma Lattes. **Ciência da Informação**, [*s.l.*], v. 34, n. 1, p. 64–77, 2005. Disponível em: https://doi.org/10.1590/S0100-19652005000100008. Acesso em: 11 set. 2023.

BAPTISTA, A. A.; COSTA, S. M. D. S.; KURAMOTO, H.; RODRIGUES, E. Comunicação científica: o papel da open archives initiative no contexto do acesso livre. **Encontros Bibli:** revista eletrônica de biblioteconomia e ciência da informação, Slorianópolis, SC, p. 1–17, 13 dez. 2007. Disponível em: https://doi.org/10.5007/1518-2924.2007v12nesp1p1. Acesso em: 11 set. 2023.

BRAUNER, D. F.; ARAÚJO, R. M.; SANTOS, G. R. M. Alinhamento de nomes de coautores de publicações científicas: uma abordagem prática. **International journal of knowledge engineering and management**. Florianópolis, SC. Vol. 5, n. 13 (nov. 2016/fev. 2017), p. 42-57 Disponível em: https://lume.ufrgs.br/handle/10183/169161. Acesso em: 16 ago. 2021.

CHEN, P. P.-S. The entity-relationship model—toward a unified view of data. **Transações ACM em sistemas de banco de dados,** [*s.l.*], v. 1, n. 1, p. 9–36, 1976.

COMARELA, G.; FRANCO, G.; TROIS, C.; LIBERATO, A.; MARTINELLO, M.; CORRÊA, J. H.; VILLAÇA, R. Introdução à Ciência de Dados: Uma Visão Pragmática utilizando Python, Aplicações e Oportunidades em Redes de Computadores. **Sociedade Brasileira de Computação**, Brasília, 2019. Disponível em: https://sol.sbc.org.br/livros/index.php/sbc/catalog/view/65/289/538-1. Acesso em: 11 set. 2023.

GRÁCIO, M. C. C. Colaboração científica: indicadores relacionais de coautoria. **Brazilian Journal of Information Science:** research trends, [*s.l.*], v. 12, n. 2, 2018. Disponível em: https://revistas.marilia.unesp.br/index.php/bjis/article/view/7976. Acesso em: 11 set. 2023.

KATZ, J. S.; MARTIN, B. R. What is research collaboration? **Research Policy**, v. 26, n. 1, p. 1–18, 1 mar. 1997. Disponível em: https://doi.org/10.1016/S0048-7333(96)00917-1. Acesso em: 11 set. 2023.

LEE, D.; ON, B.-W.; KANG, J.; PARK, S. Effective and scalable solutions for mixed and split citation problems in digital libraries. *In*: IQIS05: INTERNATIONAL WORKSHOP ON INFORMATION QUALITY IN INFORMATION SYSTEMS, 2005, 17 jun. 2005. **Proceedings of the 2nd international workshop on Information quality in information systems** [...]. Baltimore Maryland: ACM, 2005. p. 69–76. Disponível em: https://dl.acm.org/doi/10.1145/1077501.1077514. Acesso em: 11 set. 2023.

MIRANDA, Antonio. O Campo da ciência da informação: gênese, conexões e especialidades. *In*: AQUINO, Mirian de Albuquerque (org.). **A ciência da informação e a teoria do conhecimento objetivo**: um mal necessário. João Pessoa: Editora Universitária/UFPB, 2002. p. 9–24.

MUELLER, S. Literatura científica, comunicação científica e ciência da informação. **Para entender a ciência da informação**. Salvador: EDUFBA, 2007.

MUGNAINI, R.; DIGIAMPIETRI, L. A.; OLIVEIRA, L. C. de; FERREIRA, S. M. S. P. Normalização de nomes de autores em fontes de informação institucionais: proposta de um método automático de verificação de erros. **Em Questão**, Rio Grande do Sul, v. 18, n. 3, p. 263–279, 2012. Disponível em: https://seer.ufrgs.br/EmQuestao/article/view/33265. Acesso em: 11 set. 2023.

PINHEIRO, L. V. R.; LOUREIRO, J. M. M. Traçados e limites da ciência da informação. **Ciência da Informação**, [*s.l.*], v. 24, n. 1, 1995. Disponível em: https://revista.ibict.br/ciinf/article/view/609. Acesso em: 11 set. 2023.

PRESS, G. A Very Short History Of Data Science. [*s. d.*]. **Forbes**. Disponível em: https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/. Acesso em: 11 set. 2023.

REIS, M. de J. Ciência de dados e ciência da informação: guia para alfabetização de dados para bibliotecários. 12 jul. 2019. Disponível em: https://ri.ufs.br/jspui/handle/riufs/12667. Acesso em: 11 set. 2023.

ROBREDO, J. Da Ciência da Informação. **Revisitada aos Sistemas Humanos de Informação**. Brasília: Thesaurus, 2003.

ROLIM, M. V. Análise do perfil do profissional da informação para a atuação como cientista de dados em ambientes de big data: uma perspectiva a partir das disciplinas do curso de Biblioteconomia da UnB. 3 jul. 2018. Disponível em: https://bdm.unb.br/handle/10483/20898. Acesso em: 11 set. 2023.

SAMPIERI, R. H.; COLLADO, C. F.; LUCIO, M. D. P. B.; MORAES, D. V. de; GARCIA, A. G. Q.; JÚLIO, M.; SILVA, D. da. **Metodologia de Pesquisa**. 5a edição. [*s. l.*]: Penso, 2013.

SARACEVIC, T., Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, Minas Gerais, v. 1, n. 1, 1992. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/235. Acesso em: 27 set. 2021.

TARGINO, M. das G. Comunicação científica: uma revisão de seus elementos básicos. **Informação & Sociedade:** Estudos, Paraíba, 2000. Disponível em: https://periodicos.ufpb.br/ojs/index.php/ies/article/view/326. Acesso em: 17 ago. 2021.

VALEIRO, P. M.; PINHEIRO, L. V. R. Da comunicação científica à divulgação. **Transinformação**, Campinas, v. 20, p. 159–169, 2008. Disponível em: https://brapci.inf.br/index.php/res/v/116012. Acesso em: 25 out. 2021.

VICTORINO, M. de C. **Organização da Informação para dar Suporte à Arquitetura Orientada a Serviços: Reuso da Informação nas Organizações**. 2011. 280 f. Tese de doutorado – Universidade de Brasília, Brasília, 2011.

WELLISCH, H. H. **A cibernética do controle bibliográfico**: para uma teoria dos sistemas de recuperação da informação. Brasília: IBICT, 1987.