UNIVERSIDADE ESTADUAL PAULISTA – UNESP
Campus de Marília
Faculdade de Filosofia e Ciências
Programa de Pós-Graduação em Ciência da Informação

MODELO PARA AUTOMATIZAÇÃO DO PROCESSO DE DETECÇÃO E EXTRAÇÃO DE REFERÊNCIAS BIBLIOGRÁFICAS

FÁBIO HENRIQUE ALVES

Marília
2022

UNIVERSIDADE ESTADUAL PAULISTA – UNESP
Campus de Marília
Faculdade de Filosofia e Ciências
Programa de Pós-Graduação em Ciência da Informação

FÁBIO HENRIQUE ALVES

MODELO PARA AUTOMATIZAÇÃO DO PROCESSO DE DETECÇÃO E EXTRAÇÃO DE REFERÊNCIAS BIBLIOGRÁFICAS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Informação, da Faculdade de Filosofia e Ciências, da Universidade Estadual Paulista, Campus de Marília, como um dos requisitos para a obtenção do título de Mestre em Ciência da Informação.

Orientador: Prof. Dr. José Eduardo Santarém Segundo

Área de Concentração: Informação, Tecnologia e Conhecimento

Linha de Pesquisa: Informação e Tecnologia

Marília
2022

FÁBIO HENRIQUE ALVES

MODELO PARA AUTOMATIZAÇÃO DO PROCESSO DE DETECÇÃO E EXTRAÇÃO DE REFERÊNCIAS BIBLIOGRÁFICAS

BANCA EXAMINADORA:

Prof. Dr. Jose Eduardo Santarém Segundo (Orientador)
Universidade de São Paulo (USP) – Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto-SP/Universidade Estadual Paulista (UNESP) – Faculdade de Filosofia e Ciências de Marília-SP.

Prof. Dr. Leonardo Castro Botega (Membro Titular Interno)
Universidade Estadual Paulista (UNESP) – Faculdade de Filosofia e Ciências de Marília-SP.

Prof. Dr. João de Melo Maricato (Membro Titular Externo)
Universidade de Brasília (UnB) – Faculdade de Ciências da Informação

Marília, 2 de agosto de 2022.

A474m

Alves, Fábio Henrique

Modelo para automatização do processo de detecção e extração de referências bibliográficas / Fábio Henrique Alves. -- Marília, 2022

96 f. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Faculdade de Filosofia e Ciências, Marília

Orientador: José Eduardo Santarém Segundo

1. Referências. 2. Citação. 3. Comunicação científica. 4. Extração de referências. 5. Identificação de referências. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Filosofia e Ciências, Marília. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

(...) Vejo a vida passar num instante

Será tempo o bastante que tenho pra viver?

Não sei, não posso saber

Quem segura o dia de amanhã na mão?

Não há quem possa acrescentar um milímetro a cada estação

Então, será tudo em vão? Banal? Sem razão?

Seria. Sim, seria, se não fosse o amor

O amor cuida com carinho

Respira o outro, cria o elo

O vínculo de todas as cores

Dizem que o amor é **amar-elo…**

Canção Principia

(Emicida e Henrique Vieira)

Dedico à minha mãe Clarice e ao meu pai José, que com muita dedicação, diretamente ou indiretamente, me fizeram acreditar que a busca pela educação seria o melhor caminho a ser seguido por mim.

AGRADECIMENTOS

Agradeço primeiramente a Deus e a todos meus anjos e santos protetores por terem me ajudado no decorrer dessa jornada acadêmica e científica.

À minha esposa Beatriz Rosa Pinheiro Alves, por ter me apoiado, me motivado e ter sido a melhor companheira pessoal e profissional que eu poderia ter tido no decorrer desse mestrado. Que sorte ter você por perto todos os dias da minha vida, meu amor. Com certeza, você contribuiu para que eu percebesse o mundo de maneira diferente, e principalmente, a perceber o quanto a ciência e pesquisa são essenciais para construção de um mundo no qual eu acredito que podemos alcançar, e hoje eu faço parto desse movimento. Obrigado, obrigado e obrigado!

Agradeço à minha mãe Clarice, pai José e às minhas irmãs Josimara e Patrícia. Vocês são minha base sólida, em que me apoio e me percebo enquanto ser humano seguro de que sou amado e de que posso amar.

A todos os professores do Programa de Pós-Graduação em Ciência da Informação da Unesp de Marília pelos conhecimentos compartilhados.

Ao meu orientador Prof. Dr. José Eduardo Santarém Segundo, pela humanidade, respeito, paciência e muito conhecimento compartilhado no decorrer desta pesquisa. Foi uma honra ter sido seu orientando nesta jornada tão importante na minha vida. Muito obrigado por todo apoio.

Por fim, meus agradecimentos a todos que não foram citados, mas que certamente contribuíram diretamente ou indiretamente para que esta pesquisa se concretizasse. Meu muito obrigado!

ALVES, Fabio Henrique. **Modelo para automatização do processo de detecção e extração de referências bibliográficas**. Orientador: Jose Eduardo Santarém Segundo. 2022. 94f. Dissertação (Mestrado em Ciência da Informação) – Programa de Pós-Graduação em Ciência da Informação – Faculdade de Filosofia e Ciências – Universidade Estadual Paulista (UNESP), Marília, 2022.

RESUMO

Os estudos de citação são fundamentais para o desenvolvimento da ciência, em que a multidisciplinaridade se faz presente em grande parte das pesquisas científicas, tornando a citação e referência, fatores primordiais no domínio da comunicação científica. Este trabalho contextualiza uma estrutura de rede de informações originada por dados presentes em produção científica, tais como autores, citações, relações entre os autores e as obras científicas, objetivando constituir um modelo de processamento de informações que torne possível identificar, extrair e organizar esses dados de textos em língua portuguesa, baseados nas normas ABNT. Para isso, utilizou-se uma metodologia de natureza qualiquantitativa, em que se criou um modelo utilizando conceitos de citação e referência e posteriormente, foi implementado um protótipo como prova de conceito do modelo elaborado. Diante disso, o estudo parte do seguinte problema: como é possível criar indicadores ou realizar análises de citações e referências bibliográficas em segmentos gerais ou específicos, sem depender de outras entidades, como editoras e jornais. Portanto, espera-se que, com o resultado dessa pesquisa seja possível propor uma solução que seja capaz de identificar, extrair e organizar referências bibliográficas e citações de documentos científicos em qualquer segmento, gerando contribuições significativas para apoiar o desenvolvimento das análises de bases de dados científicas e a construção de novas soluções direcionadas para a ciência, propondo um vínculo aprofundado sobre os dados presentes na pesquisa.

Palavras-chave: Referências; Citação; Comunicação científica; Extração de referências; Identificação de referências.

ALVES, Fabio Henrique. **Automated model for identification and identification of identification processes.** Advisor: José Eduardo Santarém Segundo. 2022. 94f. Dissertation (Master in Information Science) – Postgraduate Program in Information Science – Faculty of Philosophy and Sciences – Universidade Estadual Paulista (UNESP), Marília, 2022.

ABSTRACT

Citation studies are fundamental for the development of science, in which multidisciplinarity is present in most scientific research, making citation and reference key factors in the field of scientific communication. This work contextualizes an information network structure originated by data present in scientific production, such as authors, citations, relationships between authors and scientific works, aiming to constitute an information processing model that makes it possible to identify, extract and organize these data. of texts in Portuguese, based on ABNT standards. For this, a qualitative-quantitative methodology was used, in which a model was created using citation and reference concepts and later, a prototype was implemented as proof of concept of the elaborated model. Therefore, the study starts from the following problem: how is it possible to create indicators or carry out analyzes of citations and bibliographic references in general or specific segments, without depending on other entities, such as publishers and newspapers. Therefore, it is expected that, with the result of this research, it will be possible to propose a solution that is capable of identifying, extracting and organizing bibliographic references and citations of scientific documents in any segment, generating significant contributions to support the development of database analysis. and the construction of new solutions aimed at science, proposing a deep link with the data present in the research.

Keywords: References; Citation; Scientific communication; References extraction; References identification.

LISTA DE FIGURAS

P.

LISTA DE QUADROS

P.

LISTA DE TABELAS

P.

SUMÁRIO

1 INTRODUÇÃO

Os estudos de citação são realizados pelos pesquisadores como um fator fundamental para o desenvolvimento da ciência. Diante da multidisciplinaridade envolvida em grande parte das pesquisas, os atos de citação e referência contribuem de maneira segura e significativa com a causa dos cientistas, no domínio da Comunicação Científica. De acordo com Grácio (2016), a análise relacional das citações possibilita o conhecimento acerca das relações estruturais e sociais que se estabelecem entre documentos e pesquisadores. Hjorland (2013) aponta que a análise de citações contribui para identificação das características peculiares a abrangentes dos pesquisadores, suas publicações e seus pesquisados, bem como os fatores de impacto e nível de domínio de tais elementos. Portanto, reconhecer contribuições de distintos campos e domínios é um contrato que rege termos legais e éticos, fornecendo fácil acesso às outras áreas de pesquisa e facilitando a interoperabilidade com outras temáticas.

A literatura especializada apresenta duas teorias em relação aos estudos de citação, a primeira, denominada teoria normativa, que é amparada na ideologia de que a ciência é um campo governado por práticas de citação decorrentes de dívidas intelectuais, recompensas e sanções internas, ou seja, livre de interferências sociais e culturais externas (LEYDESDORFF, 1998; NICOLAISEN, 2007). A segunda teoria, chamada de teoria construtivista, possui um foco mais direcionado para identificar os motivos que subsidiam as origens das citações realizadas pelos cientistas, com o propósito de tentar estabelecer a trajetória percorrida para a construção do conhecimento.

A Comunicação Científica é evidenciada por meio de diversas práticas no âmbito científico. Portanto, Leydersdorff e Wouters (1999) destacam que os contextos sociais e culturais estão diretamente envolvidos no processo de citação, que contribuem positivamente através de uma formalização entre a produção, citação e referência. No entanto, esse processo pode ser utilizado como estratégia de manipulação, convencimento ou em situações em que os pesquisadores ignoram determinadas regras para enriquecer suas contribuições.

Com base nas citações supracitadas, torna-se possível observar que as referências bibliográficas estão relacionadas não somente à pesquisa em que se faz aplicação, mas também com um universo científico presente na Web, criando-se uma conexão multidisciplinar; com possibilidades de realizar a interoperabilidade entre as mais diversas áreas do conhecimento. Diante desses aspectos, com a grande quantidade de recursos presentes na Web, a tecnologia pode contribuir significativamente com o desenvolvimento de soluções para melhorar esses processos, uma vez que auxilia no método de citação e referência, a partir da disponibilização de ferramentas computacionais para apoio ao processo de inserção, alteração ou exclusão de uma determinada citação presente em uma obra científica, por meio de programas de licença livre ou privada.

Os seres humanos possuem a capacidade de interpretar dados e formular informações para se obter conhecimento sobre um determinado conceito. No entanto, os sistemas computacionais não são capazes de interpretar conceitos em um determinado contexto, logo, não processam corretamente o conteúdo de um recurso informacional estipulado, ocasionando muitas vezes, a recuperação de informações de forma errônea, não correspondendo às necessidades do usuário. De acordo com Santarém-Segundo e Vidotti (2003), o ser humano é capaz de associar significados, estabelecendo novos contextos, diferentemente dos computadores, que apesar de trabalharem com processamento lógico, não são capazes de fazer associações semânticas, consequentemente, não assimilam novos conhecimentos.

Este trabalho contextualiza uma estrutura de rede de informações originada por dados presentes nos materiais científicos, tais como autores, citações, produções científicas, relações entre os autores e as obras científicas, enfim, parte do escopo que envolve as referências bibliográficas de uma pesquisa científica, objetivando constituir um modelo de processamento de informações que torne possível identificar, extrair e organizar esses dados de textos em língua portuguesa, baseados nas normas ABNT. Diante disso, o estudo parte do seguinte problema: como é possível criar indicadores quantitativos ou realizar análises de citações e referências bibliográficas em

segmentos gerais ou específicos, sem depender de outras entidades, como editoras, jornais e bancos de dados?

Diante desta perspectiva, após a construção do modelo proposto nesta pesquisa, que é composto por informações bibliográficas e suas possíveis relações com objetos internos e externos, espera-se que, com o resultado dessa pesquisa seja possível propor uma solução capaz de identificar, extrair e organizar referências bibliográficas e citações de documentos científicos em qualquer segmento, trazendo contribuições significativas para apoiar o desenvolvimento de análises de bases dos dados científicos, o desenvolvimento de novas soluções direcionadas à ciência e processos voltados para a apresentação, descrição, recuperação e organização da informação, propondo um vínculo sobre os dados presentes nas pesquisas.

1.1 PROBLEMA

O avanço da tecnologia impactou positivamente o desenvolvimento e publicação de novos trabalhos científicos, em todas as áreas do conhecimento e campos do saber. Tal avanço é justificado por inúmeras razões, entre elas, o fácil acesso aos materiais já publicados e disponibilizados nas redes de informação, ponto que facilita a busca por referências para a construção de novas pesquisas científicas.

Os trabalhos científicos possuem metadados que se associam aos dados do trabalho e dos autores. Normalmente as plataformas que hospedam os materiais científicos utilizam esses metadados para apresentarem informações relacionadas ao trabalho, como autores, quantidade de páginas, assunto, entre outras variáveis. De acordo com Alves (2010, p.47), os metadados:

> [...] são atributos que representam uma entidade (objeto do mundo real) em um sistema de informação. Em outras palavras, são elementos descritivos ou atributos referenciais codificados que representam características próprias ou atribuídas às entidades; são ainda dados que descrevem outros dados em um sistema de informação, com o intuito de identificar de forma única uma entidade (recurso informacional) para posterior recuperação (ALVES, 2010, p.47).

No cenário atual, somente grandes editoras, sustentadas por eficientes plataformas tecnológicas que englobam revistas, jornais, e todo artefato de comunicação que divulga um trabalho científico possuem os dados referentes à obra e seus respectivos autores. No entanto, essas plataformas utilizam um sistema de *input* de informações, isto é, o processo de submissão de um trabalho científico, por exemplo, geralmente requer os seguintes passos:

1. Efetuar o cadastro na plataforma que hospeda a revista ou artefato de comunicação científica;
2. Criar um registro de submissão do trabalho científico na modalidade desejada (resumo expandido, trabalho completo, entre outros);
3. Adicionar os metadados dos autores, bem como documentos, escolaridade, profissão, áreas de interesse, dentre outros.
4. Adicionar os metadados relacionados com o trabalho em questão, como número de páginas, autores mais citados, resumo do trabalho, palavras-chave, referências, entre outros.

Os metadados solicitados podem variar entre as plataformas, porém, todos possuem o mesmo objetivo, obter o máximo de dados possíveis para que isso possa servir como referência para outras aplicações tecnológicas. Entretanto, além do fato de que esses recursos podem ser fornecidos incorretamente, somente cadastrar os metadados para esses recursos informacionais pode não ser o suficiente para atingir o grau de sucesso no que tange aos processos, como a organização e a recuperação desses recursos. Para isso, os metadados devem ser implementados seguindo padrões, que normalmente são implementados pelas plataformas, garantindo a interoperabilidade entre os sistemas que irão consumir esses recursos. Apesar de todas as informações dos trabalhos submetidos, somente estão disponíveis análises em um contexto geral, o que impossibilita a geração de novas informações, tais como:

- Identificar quais foram os autores mais citados e menos citados em um período específico;

- Identificar autores mais citados e menos citados por área do conhecimento;
- Identificar e quantificar as revistas, jornais, entre outros artefatos de divulgação científica;
- Identificar a quantidade de vezes que um autor foi citado no texto;
- Identificar a parte do texto em que o(s) autor(es) foram citados.

Todavia, ressalta-se ainda que existe complexidade em disponibilizar todas essas informações, seja por necessidade de novas tecnologias, melhoria de processos ou adaptação de estruturas de armazenamento. Porém, a perspectiva conveniente é que os dados, enquanto recursos principais para a possibilidade de execução dessas atividades, estão disponíveis. Sem a presença dos dados, dificilmente será possível gerar essas informações.

Nesse contexto, a I4OC (2022) relata que o sistema atual de comunicação acadêmica expõe inadequadamente as redes de conhecimento que já existem na literatura. Para a organização, os dados de citação geralmente não ficam disponíveis gratuitamente para acesso externo, geralmente estão vinculados a licenças inconsistentes e difíceis de serem analisadas, além de se mostrarem ilegíveis por máquinas computacionais.

É importante ressaltar que, o foco desta pesquisa não está em resolver problemas voltados para questões específicas sobre as atividades abordadas anteriormente, ou seja, não cabe identificar os motivos e causas de um determinado autor ser mais ou menos citado. Portanto não há qualquer tipo de inferência acerca das questões que envolvem citação e referência, ou então qualquer tipo de abordagem bibliométrica ou cientométrica nesse contexto de estudo, mesmo que um dos objetivos dessa pesquisa seja direcionado a apoiar, agilizar e facilitar os estudos nessas referidas áreas. Com o aporte dos dados, a dificuldade para transformá-los em informações se dá em outros segmentos, como identificar as melhores ferramentas tecnológicas para tratar grandes quantidades de dados, apresentar formas de visualização de modo que tais informações possam ser utilizadas para diferentes propósitos.

Diante de tais considerações, observa-se ainda que, a ciência brasileira não está indexada adequadamente em bases de dados internacionais, o que dificulta a criação de indicadores ou na análise de processos que possuem

referências ou citações, visto que somente grandes editoras e bases de dados específicas têm capacidade para isso. Com isso, essas atividades, tais como geração de indicadores, análises de citações e referências, tornam-se dependentes de editoras ou bases de dados específicas, administradas por entidades que executam essas tarefas mediante seus critérios e interesses. A falta de indexação nas bases internacionais impacta na quantidade de estudos que analisam, por exemplo, a ocorrência de autores no texto.

Nesse contexto, Barata (2015) discorre sobre o problema da produção científica brasileira que não pode ser contabilizada, como no caso das áreas de conhecimento que tradicionalmente publicam em língua portuguesa, abordando sobre a ênfase dos indicadores de publicações científicas internacionais, ressaltando a consequência prejudicial aos pesquisadores, onde, quem não está indexado, não possui indicador. Em 2015, estimou-se que existiam cerca de mil revistas científicas no Brasil, das quais 400 estão indexadas em bases brasileiras, como a SciELO (*Scientific Eletronic Library Online*), porém, há diversas outras publicações hospedadas em outros lugares, como portais de periódicos científicos relacionados pelo Instituto Brasileiro de Informação em Ciência e Tecnologia – Ibict, que utilizam o Sistema Eletrônico de Editoração de Revistas - SEER, além de várias outras publicações que não possuem indexação (BARATA, 2015).

O *Google Scholar*, por exemplo, disponibiliza um gráfico sobre o índice de citações de um determinado autor em um contexto geral, incluindo os trabalhos citados, conforme apresenta-se.

Figura 1: Relação de autor e citação no Google Acadêmico

Fonte: Google Acadêmico (2021)

Entretanto, essas informações não englobam todas as publicações do autor em questão, visto que isso depende da forma como a plataforma extrai os dados de outras fontes. Este cenário torna-se ainda mais problemático, quando considerado em segmentos específicos, como eventos, fascículo de um periódico científico ou linhas específicas de um determinado contexto.

Geralmente, para a submissão de trabalhos em eventos científicos, a plataforma solicita o cadastro do usuário e o envio do documento científico para análise dos avaliadores e posteriormente publicação, caso seja aprovado. No caso, o evento solicita somente que o usuário faça o upload do documento, onde não há opção para que o usuário informe quais foram metadados relacionados às referências. Entretanto, quando isso ocorre, essas informações não acompanham a publicação quando disponibilizadas ao público ou então não se enquadram em um cenário factível para realizar análises ou estudos. Com isso, responder às questões sobre inferências levantadas anteriormente, pode ser uma tarefa complexa, demorada e passível de erros. Isso ocorre devido à necessidade de um trabalho manual de identificação das referências utilizadas em cada trabalho, bem como validar se essas referências estão de fato sendo usadas, pois podem ter sido inseridas erroneamente.

De acordo com I4OC (2022), uma iniciativa de citação aberta, constituída por editores acadêmicos, pesquisadores e outras partes interessadas para prover a disponibilidade irrestrita de dados de citações acadêmicas; o número

de publicações acadêmicas é estimado para dobrar a cada nove anos, o que permite que pesquisadores se mantenham a par de desenvolvimentos significativos, em diferentes campos da ciência. Porém, para que isso seja possível, é essencial ter acesso irrestrito aos dados de referências bibliográficas e citações em formato legível por máquina.

Os conceitos entorno da problemática desta pesquisa serão apresentados no capítulo 3, tais como a influência da ciência na sociedade, os conceitos e abordagens da comunicação científica na ciência e as perspectivas de citações e referências vinculadas aos cientistas.

1.2 JUSTIFICATIVA

A tecnologia assume um papel relevante no que diz respeito aos estudos de citação e referência. Com o aumento progressivo do volume de pesquisas científicas presentes na Web, as iniciativas que buscam melhorar processos como recuperação, apresentação e organização das informações adquirem uma nova perspectiva, onde a Ciência da Informação se faz presente no desenvolvimento de novas soluções.

Baseando-se nas teorias de que as referências bibliográficas de uma pesquisa científica podem fazer parte não só da pesquisa em que se aplicam, mas de um cenário global de ciência, torna-se viável o desenvolvimento de ferramentas que auxiliam esse processo, de forma a agilizar o ato de minerar as referências e organizá-las de forma automática e confiável, com o intuito de poder utilizar esses dados para outras soluções tecnológicas. Identificar as referências de uma determinada obra científica e relacioná-las com o próprio texto de forma manual é demorado, e, principalmente, passível de erros, uma vez que, o processo realizado por humanos pode conter falhas, impactando diretamente e, negativamente, na qualidade do resultado.

Diante de tais considerações, observou-se que os estudos existentes nessa área são feitos tomando como ponto de partida as referências e não o texto, em si. Logo, nota-se a necessidade de estudos que analisem a aplicabilidade de tecnologias para auxiliar o processo de mineração de referências bibliográficas e suas relações com o texto em questão, para que este processo possa servir de parâmetro em outras aplicações e estudos,

como a criação de indicadores quantitativos ou análises de citações e referências bibliográficas.

Embora toda prática requeira a necessidade de fundamentação teórica que a sustente, é notável a importância de pesquisas que possibilitem a criação de soluções informacionais, unindo teorias e práticas computacionais.

1.3 OBJETIVOS

A presente pesquisa possui como objetivo constituir um modelo de processamento de informações que torne possível identificar, extrair e organizar referências e citações de textos em língua portuguesa, baseados nas normas ABNT.

Como objetivos específicos, foram definidos:

- Compreender o funcionamento de referência e citação em textos da língua portuguesa;
- Compreender os padrões de citação e identificar seus padrões de construção;
- Definir um processo que possa identificar, organizar e quantificar referência e citação em textos que usem padrão ABNT em língua portuguesa;
- Constituir uma prova de conceito que possa demonstrar o modelo proposto.

Este trabalho é composto por seis seções, divididas entre as etapas de introdução, desenvolvimento e considerações finais. A pesquisa foi distribuída e representada pelas seções de: ‘Introdução’, ‘Metodologia, ‘Comunicação científica e seus atributos’, ‘Citação: conceitos e abordagens’, ‘Modelo para detecção e extração de referências bibliográficas’, ‘Implementação do protótipo’’ e ‘Considerações finais’.

2 METODOLOGIA

Neste capítulo, será apresentada a metodologia, sua caracterização e a trajetória metodológica para atender aos objetivos propostos desta pesquisa. Diante da ampla quantidade de métodos e metodologias disponíveis para o desenvolvimento do estudo, foram identificados e utilizados aqueles que melhores se adaptaram à situação proposta. Para Prodanov e Freitas (2013), no desenvolvimento prático de uma pesquisa, mescla-se todos os tipos de metodologia, afunilando em uma ou outra para realmente fazer uso, de modo que nenhum tipo de pesquisa é autossuficiente. Nesse contexto, de acordo com Demo (2000, p.44)

> [...] todas as pesquisas são ideológicas, pelo menos no sentido de que implicam posicionamento implícito por trás de conceitos e números; a pesquisa prática faz isso explicitamente. Todas as pesquisas carecem de fundamento teórico e metodológico e só têm a ganhar se puderem, além da estringência categorial, apontar possibilidades de intervenção ou localização concreta (DEMO, 2000, p.44).

O quadro 1 demonstra as definições e autores respectivamente mais utilizados, referente às características da pesquisa.

Quadro 1: Classificação de autores e definições como base para metodologia da pesquisa

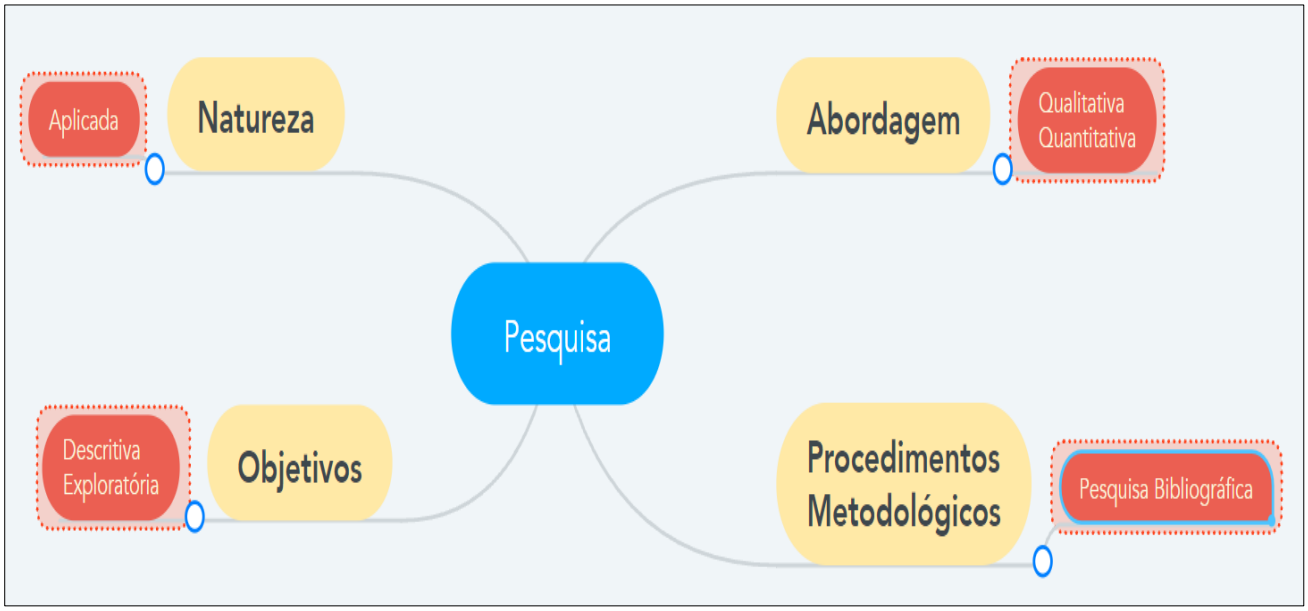
Critério	Classificação	Descrição	Autor (es)
Natureza	Aplicada	Produz conhecimentos para aplicação prática dirigidos à solução de problemas específicos, envolvendo verdades e interesses locais	Prodanov e Freitas (2013)
		Associam-se à aplicação prática de conhecimentos para resolver problemas sociais	Boaventura (2004)
		Associam-se à aplicação imediata de conhecimentos em um contexto circunstancial, relevando o desenvolvimento de teorias	Prodanov e Freitas (2013) via Gil (2008)
Objetivos	Exploratória	Conseguir novas percepções e contribuir para novos avanços sobre o objeto estudado, bem como proporcionar maior familiaridade com o problema, tornando-o mais transparente ou possibilitando a construção de hipóteses	Gil (2009)
		Proporciona maior familiaridade com o tema, elas o tornam explícito e permitem a construção de hipóteses sobre o mesmo.	Prodanov e Freitas (2013)
		Explorar é a primeira aproximação de um tema e visa criar maior familiaridade em relação a um fato ou fenômeno, de modo que possam informar ao pesquisador a real importância do problema, o estágio atual da situação e possivelmente revelar novas fontes de informação	Santos (2015)
	Descritiva	Registra, analisa e ordena dados, sem interferir no resultado, além de classificar, explicar e interpretar as rotinas de execução, expondo características de uma determinada população ou fenômenos, que demandam técnicas padronizadas de coleta de dados	Prodanov e Freitas (2013)
Abordagem	Qualitativa	O ambiente natural ser uma fonte direta para coleta de dados, além de um ambiente propício para interpretar fenômenos e atribuir significados	Prodanov e Freitas (2013)
	Quantitativa	Utiliza-se recursos e técnicas de estatística, procurando gerar conhecimentos a partir dos números obtidos pelo pesquisador.	Prodanov e Freitas (2013)

Fonte: Elaborado pelo autor (2021).

2.1 CARACTERÍSTICAS DA PESQUISA

A presente pesquisa é de natureza aplicada, utilizando-se de uma abordagem qualiquantitativa, caracterizada como descritiva e exploratória. Possui como universo de pesquisa os aspectos epistemológicos voltados para o escopo de citações e referências bibliográficas, especificamente nas atividades presentes no processo de referenciar outras pesquisas. A figura 2 demonstra a caracterização da pesquisa.

Figura 2: Caracterização da pesquisa



Fonte: Elaborado pelo autor (2021).

Em relação à natureza da pesquisa, define-se pela natureza aplicada, onde busca-se gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos, envolvendo verdades e interesses locais (Prodanov e Freitas, 2013). Boaventura (2004), aborda que pesquisas de natureza aplicada podem estar associadas a aplicação prática de conhecimentos para resolver problemas sociais. Complementando o pensamento de Boaventura (2004), Prodanov e Freitas (2013) relatam que as pesquisas aplicadas possuem ligação com a aplicação imediata de conhecimentos em um contexto circunstancial, relevando o desenvolvimento de teorias. Caracteriza-se também, segundo os procedimentos de coleta, como pesquisa de experimento, ocorrendo quando um fato ou fenômeno da realidade é reproduzido de forma

controlada, objetivando descobrir os fatores que o produzem ou a razão pelas quais são produzidos (Santos, 2015).

Optou-se pelo caráter exploratório devido à necessidade de conseguir novas percepções e contribuir para novos avanços sobre o objeto estudado, bem como proporcionar maior familiaridade com o problema, tornando-o mais transparente ou possibilitando a construção de hipóteses (GIL, 2009). Prodanov e Freitas (2013) relatam que além das pesquisas com caráter exploratório proporcionarem maior familiaridade com o tema, elas o tornam explícito e permitem a construção de hipóteses sobre o mesmo. Santos (2015) corrobora com Gil (2009), ressaltando que o ato de explorar é a primeira aproximação de um tema e visa criar maior familiaridade em relação a um fato ou fenômeno, de modo que possam informar ao pesquisador a real importância do problema, o estágio atual da situação e possivelmente revelar novas fontes de informação.

Configura-se também, como pesquisa descritiva, pois o objeto de estudo foi identificado através da extração de conceitos presentes na literatura e, de acordo com Prodanov e Freitas (2013), uma pesquisa com objetivos descritivos registra, analisa e ordena dados, sem interferir no resultado, além de classificar, explicar e interpretar as rotinas de execução, expondo características de uma determinada população ou fenômenos, que demandam técnicas padronizadas de coleta de dados. Logo, o modelo para extração e organização dos recursos bibliográficos será implementado e utilizado para como prova para suprir as necessidades apresentadas no projeto diante de uma perspectiva de enriquecimento semântico.

Quanto à abordagem, a pesquisa utiliza métodos qualitativos e quantitativos. Para Prodanov e Freitas (2013), abordagem qualitativa diz respeito ao ambiente natural ser uma fonte direta para coleta de dados, além de um ambiente propício para interpretar fenômenos e atribuir significados. Em relação à abordagem quantitativa, há a necessidade de utilizar recursos e técnicas de estatística, procurando gerar conhecimentos a partir dos números obtidos pelo pesquisador.

No que se refere à abordagem quantitativa, Silveira e Caregnato (2017) relatam que de modo geral, a perspectiva orientada por estudos de citação contempla as citações e referências como objetos de estudos, que podem ser utilizados em diversos campos e domínios da ciência em conjunto com

métodos e técnicas de base quantitativa. Entretanto, o enfoque maior da pesquisa não se dá em analisar os dados obtidos por meio da solução implementada, mas sim em disponibilizar uma solução viável para tal atividade.

A pesquisa foi constituída por três etapas: (i) revisão teórica sobre os conceitos e abordagens envolvidos no estudo, que diretamente irão contribuir para a construção do modelo e protótipo para extração e detecção de referências bibliográficas; (ii) Elaboração do modelo que seja capaz de identificar, extrair e organizar as referências bibliográficas de trabalhos acadêmicos; e iii) Implementação do protótipo como prova de conceito embasado modelo elaborado.

De acordo com Gil (2009), a revisão bibliográfica se pauta pela busca, exploração e análise de diversos tipos de materiais bibliográficos que se relacionam com o tema da pesquisa, como artigos científicos, teses, dissertações e livros. O autor enfatiza que a pesquisa bibliográfica é importante para o pesquisador atingir seus objetivos e responder de maneira mais efetiva os problemas de pesquisa. Além disso, é por meio do referencial teórico que se torna possível alinhar e discutir os resultados com cientificidade.

Para a presente pesquisa, o objetivo principal é embasado em um pensamento sistêmico, estudado a partir da Teoria Geral dos Sistemas, de Bertalanffy e Boulding, que objetivaram desenvolver uma metodologia cuja finalidade era resolver problemas que a metodologia analítica não fosse capaz de tratar (MARTINELLI; VENTURA, 2005).

Ressalta-se que o modelo, bem como o protótipo que se pretende desenvolver nesta pesquisa, a fim de identificar, extrair e organizar referências bibliográficas poderá ser utilizado em outros cenários, pois as informações estarão disponíveis, de forma tratada, e consequentemente, com maior consistência, ou seja, a ideia é contribuir para que esse ecossistema de inovação possa ser articulado de modo que todos consigam obter resultados significativos, independentemente do tempo de sua experiência, e essa é uma é uma das soluções que esse modelo pretende difundir.

Acredita-se que essa trajetória metodológica possa ser suficiente para o alcance dos objetivos propostos e para a resposta do problema central da pesquisa.

2.2 TRABALHOS RELACIONADOS

Para a localização dos trabalhos relacionados, foram considerados os seguintes conjuntos de termos na língua portuguesa e inglesa, no Google Scholar.

- Extração de referências;
- *Reference extract;*
- *Reference parsing;*
- *Identify and extract references.*

Como resultado, os termos apareceram separadamente em alguns casos, ou seja, os resultados consideraram qualquer trabalho que incluísse algum dos termos em seu conteúdo. Diante disso, foram considerados aqueles que apareceram pelo menos dois termos no título do trabalho.

O primeiro trabalho analisado foi desenvolvido por Mattia Chenet, em 2017, intitulado como "*Identify and extract entities from bibliography references in a free text"*. Neste trabalho, foi apresentado um processo de extração e classificação de entidades de um texto não estruturado. Para isso, foi considerado um banco de dados denominado Scopus, administrado pela Elsevier, uma editora científica. Neste banco de dados, são armazenados artigos científicos, livros e manuscritos. Além dos materiais científicos, são armazenados também os perfis dos autores vinculados às suas publicações e à quantidade de vezes que o autor foi citado em outros trabalhos.

De acordo com Chenet (2017), em algumas situações, o número de citações ou de documentos vinculados a um determinado autor pode estar incorreto, isso pode ocorrer, principalmente, quando o documento está fora do padrão e, consequentemente, não foi referenciado no banco de dados ou quando o trabalho ainda está em fase de produção, e ainda não foi publicado. Quando isso ocorre, o autor deve enviar um e-mail para o suporte com as referências do material científico faltante, para que seja feita a inclusão do documento científico manualmente. No e-mail enviado ao suporte, em alguns casos, dados primordiais não são enviados, como ano de publicação, título, autores, etc. O atendimento manual prestado pelo suporte da editora pode demorar um tempo considerável, podendo chegar a meses de espera pela correção. A pesquisa procura solucionar este tipo de falha, buscando reconhecer e automatizar o padrão de documentos requerido pela editora.

Dando continuidade ao trabalho desenvolvido por Chenet (2017), inicialmente, o sistema deve reconhecer se o documento está contido em texto não estruturado ou informal, como e-mails. Para isso, o e-mail foi definido como parâmetro para treinar o sistema, pois, sua forma de escrita não requer convenções que definem sua escrita. Logo, foi utilizada uma amostra de 74.000 e-mails tratados pelo time de suporte em 2016. Com isso, observou-se que, caso fosse encontrado um ano no e-mail, as chances de existir uma referência eram maiores, logo, foi implementada uma expressão regular para obter os caracteres que correspondem a um ano, que, nos casos em que havia ocorrência, o e-mail era considerado, senão, era descartado. Com base nesta regra, cerca de 20.000 e-mails aderiram ao padrão especificado, com isso, a pesquisa selecionou 1000 registros de maneira aleatória.

Ainda sobre o trabalho de Chenet (2017), o próximo passo foi feito com base em uma ferramenta denominada *Brat,* que possui licença de código aberto, conceitos web para sua construção e pode ser utilizada para anotação de texto, que inclui funcionalidades como reconhecimento de entidades, como nome, sintaxe de dependência e verbos. Esta tecnologia é suportada por mecanismos de processamento de linguagem natural. Esta tarefa demanda um tempo maior, pois requer precisão e curadoria de dados para treinamento de um determinado algoritmo de aprendizado de máquina. Com a análise realizada pela ferramenta *Brat*, obteve-se como resultado, uma lista de entidades anotadas, tais como:

- Título – título do documento científico;
- Autores – nome dos autores que desenvolveram a pesquisa;
- Jornal/Revista – jornal onde o material científico foi publicado;
- Volume – o número do volume no meio publicado;
- Ano – ano de publicação do material científico;
- Doi – o identificador digital do material publicado;
- Páginas – número de páginas do material científico;
- Editor – nome do editor;
- ISSN – número de série que identifica o título de uma publicação.

As etapas seguintes da solução desenvolvida pelo autor se assemelham com a presente pesquisa, onde, primeiramente, identificam-se as partes de um texto que se referem a um material bibliográfico, por meio de um processo

denominado tokenização. Posteriormente, deve-se identificar os componentes de um texto, como nome dos autores, título, ano de publicação, nome do periódico, etc. Por fim, faz-se a normalização e combinação dos dados extraídos com o documento de referência.

Para tanto, foram utilizadas tecnologias voltadas para a inteligência artificial, mais especificamente o aprendizado de máquina, que, segundo Alpaydin (2014), um sistema inteligente adapta-se às mudanças, onde não há necessidade de propor soluções para todos os cenários, logo, o contexto da pesquisa feita por Chenet (2017) se adapta perfeitamente ao que se propõe esta abordagem. Por meio de rotinas como tokenização, combinação, divisão e outras, concluiu-se que é possível retirar entidades, mais especificamente referências e atributos bibliográficos de um documento não estruturado.

Zou, Le e Thoma (2010) desenvolveram uma pesquisa abordando a importância e a prática da localização e análise de artigos médicos em HTML. Para os autores, o conjunto de referências que aparece no final dos artigos científicos pode ser, em alguns casos, apenas um campo no banco de dados e, apesar de não possuírem uma estrutura mais robusta para ser armazenada, tais referências podem ser úteis para etapas de pré-processamentos em extrações automatizadas de dados bibliográficos de artigos, ou até mesmo na indexação manual ou automática de artigos. Diante da necessidade de se extrair os componentes das referências bibliográficas, como nomes dos autores, título do artigo, nome do periódico, entre outras entidades, os autores descreveram o processo de duas etapas com a utilização de dois algoritmos estatísticos de aprendizagem de máquina. O primeiro algoritmo se baseia em estatísticas de sequência e treina um campo aleatório condicional. No segundo algoritmo, são utilizadas estatísticas de recursos locais para treinar uma máquina de vetores de suporte, acompanhada por outro algoritmo que corrige sistematicamente um determinado dado caso este viole as regras predefinidas e ambos os algoritmos mantiveram desempenho similar.

Zhang, Cao e Yu (2011) elaboraram um estudo sobre citações e seu uso ubíquo em artigos biomédicos, onde representam a estrutura retórica e o conteúdo semântico dos materiais científicos. Zhang, Cao e Yu (2011) relatam sobre a complexidade da análise de citações devido aos diferentes formatos que estão enraizados em requisitos de editores ou formatos não padronizados introduzidos pelos autores. Ademais, os autores reforçam que, embora a

análise de citações não seja nova e existirem ferramentas para esta atividade, há poucas ferramentas disponíveis publicamente e poucos trabalhos abordaram análise de citações na literatura biomédica.

Para a composição do estudo, inicialmente os autores analisaram a estrutura da citação, de modo a definir exatamente quais seriam os dados extraídos. Posteriormente, foram criados os conjuntos de dados, que serviram como base de treinamento e teste para aplicar o aprendizado de máquina supervisionado. Os algoritmos de aprendizado de máquina utilizados foram supervisionados pelos Campos Aleatórios Condicionais (CRFs) para analisar automaticamente os atributos de uma citação, como autor, título, revista e ano.

Com base em um subconjunto de artigos no formato html, o estudo resultou em um pacote de código aberto que extrai automaticamente o conteúdo de uma citação, como autor, título, periódico e ano. Com a utilização de algoritmos embasados em aprendizado de máquina supervisionados, os autores obtiveram uma pontuação de 97,95%, de modo que, esta porcentagem justifica-se por meio de padrões não previstos pela solução implementada.

Os autores consideraram também o trabalho da Crossref, sendo uma organização cooperativa independente sem fins lucrativos, fundada no ano 2000 e administrada por editores científicos que objetivaram melhorar as comunicações acadêmicas por meio de infraestrutura aberta, tecnologias, ferramentas e serviços, parâmetros esses, que possuem um único objetivo: inserir o conteúdo acadêmico em contexto (CROSSREF, 2018). O Crossref possui uma plataforma ampla que oferta serviços com base em metadados disponibilizados por outras plataformas, ou seja, os resultados dependem de informações fornecidas pelos usuários nos processos de submissão de trabalhos científicos.

Álvarez (2007) elaborou um estudo com base em uma ferramenta denominada FIP (Ferramenta Inteligente de Apoio à Pesquisa) para recuperar, organizar e minerar dados de grandes coleções de documentos. Para isso, foram utilizadas diversas técnicas de identificação e extração de dados em documentos não estruturados, com o intuito de facilitar o uso dessas informações. O sistema proposto pelo autor possui uma abordagem baseada em indução de regras de etiquetagem, onde inicialmente se identifica e extrai informações presentes no corpo do artigo, como título, autores, afiliação,

resumo, entre outros elementos. Posteriormente, extrai informações presentes nas referências bibliográficas e as apresenta em formato XML.

Por meio dos trabalhos apresentados, tem-se as contribuições da presente pesquisa para a Ciência da Informação. Ressalta-se que a definição do problema em cada trabalho possui etapas diferentes, de modo que a temática e a metodologia podem ser parecidas, porém, em situações práticas distintas. Para a presente pesquisa, foi considerado o padrão de citação e referência ABNT.

De modo geral, há esforços no que tange a extrair entidades de referências bibliográficas. Alguns trabalhos visam performance, pois foram pensados com base na grande quantidade de materiais que será processado, outras soluções, porém, visam facilitar o processo, com soluções mais adaptáveis e com melhor manutenibilidade. Por meio de técnicas heurísticas e/ou conceitos de inteligência artificial, esses trabalhos visam contribuir cientificamente na construção de melhores bibliotecas digitais.

3. COMUNICAÇÃO CIENTÍFICA E SEUS ATRIBUTOS

Para o desenvolvimento desta seção, foram considerados conceitos sobre a ciência e sociedade para subsidiar teoricamente o subcapítulo de comunicação científica, bem como o capítulo de citação.

3.1 A ciência e a sociedade

A ciência se faz presente nas mais diversas atividades realizadas pela sociedade, sejam elas para fins de saúde, tecnologia, educação, biologia, entre outras áreas do conhecimento, fato que requer o reconhecimento da relevância da informação científica, do conhecimento científico, da comunidade científica e consequentemente, da comunicação científica.

A Ciência da Informação, abordada por Le Coadic (1996) como uma ciência rigorosa que faz uso de tecnologias rigorosas estuda as propriedades gerais da informação, como natureza, gênese e seus efeitos, originando mais precisamente duas frentes: (i) a análise de processos voltados à construção, comunicação e uso da informação e (ii) o desenvolvimento de produtos e serviços que permitem sua construção, a comunicação, armazenamento e uso da informação.

Ainda no que se refere à Ciência da Informação, Le Coadic (1996) sintetiza alguns temas que possuem conceitos específicos à sua origem, mas que são fortemente apoiados pela Ciência da Informação, entre eles, estão:

- campo filosófico, epistemológico, históricos, etc., de forma a abordar todo o histórico de um determinado tema, bem como sua origem, fenômenos e atributos internos;
- eletrônicos e telecomunicações, como redes, correios eletrônicos, videotexto, etc.;
- áreas econômicas, jurídicas e políticas, pois abordam a interação com a informação, como a divulgação e comercialização, de modo a envolver direitos digitais, indústrias da informação e a sociedade da informação;
- disciplinas voltadas à matemática, lógica e estatística, como algoritmos, lógica booleana e difusa, etc.;

- informática, como bases de dados, recuperação, sistemas específicos, programas para hipertexto, experiência do usuário, segurança da informação, etc.;
- sociológicos, bem como a sociologia da ciência, produtividade e comunidade científica, mérito, reconhecimento, etc.;
- psicológicos, incluindo processos heurísticos, comportamentos e representação do conhecimento, etc.;

O universo do qual faz uso da ciência não é um local isolado onde se faz presente somente o desenvolvimento de pesquisas e a comunicação destas. A ciência é um sistema social englobada por outro sistema social maior ainda, sensível a contextos políticos, sociais ou econômicos. A comunidade científica interage com as necessidades que a rodeia, sofrendo os impactos da sociedade em geral e, simultaneamente, contribuindo positivamente ou negativamente com essa sociedade (HERNANDES-CAFÍADAS, 1987).

Targino (2000) discorre sobre as responsabilidades da ciência, entre elas, a busca por desvendar e compreender a natureza e seus fenômenos por meio de métodos sistemáticos e seguros é parâmetro para obter resultados que não possuem caráter permanente, tornando a ciência uma instituição social e dinâmica, contribuindo com a evolução da humanidade, de forma a criar e modificar hábitos, publicando leis, provocando acontecimentos e ampliando as fronteiras do conhecimento.

Diante da grande quantidade de informações disponibilizadas diariamente nas redes de comunicação, muitas vezes, trabalhos mais antigos são consultados para o desenvolvimento de novas pesquisas, de forma mais atualizada e aderente à realidade atual, o que não descarta a magnitude do trabalho publicado em outras épocas, conforme aborda o autor supracitado (2000) sobre a ciência, como sendo contínua e cumulativa.

Nesse contexto, Lins de Barros (2001) relata o impacto de novas tecnologias na sociedade, como o poder computacional, a tecnologia robótica e a nanotecnologia, sendo ferramentas principais na linha de frente da explosão informacional, onde não há precedentes na história de tamanha quantidade de dados armazenada e transmitida como atualmente, fato que tende ao aumento progressivo.

A explosão bibliográfica, fenômeno comum nas mais diversas áreas do conhecimento, pode ser interpretada como o crescente número de documentos científicos produzidos e disponibilizados nas redes de comunicação. No fim do século XVII, o estabelecimento da ciência moderna e o início da publicação dos primeiros periódicos desencadeou o crescimento exponencial deste fenômeno (SOLLA PRICE, 1963).

Diante disso, a ampla quantidade de trabalhos científicos publicados nas mais diversas áreas do conhecimento contribui com o avanço da ciência, bem como da sociedade. Diante de tal perspectiva, Le Coadic (1996) sintetiza as razões que podem influenciar no crescimento da ciência:

- Ampliação dos setores relacionados à informação e comunicação, disponibilizando tecnologias para divulgação, armazenamento e recuperação do conhecimento;
- Alteração na geografia das disciplinas científicas, bem como a interdisciplinaridade e a contribuição de novas áreas para chegar a objetivos em comum;
- Surgimento de novos produtos, soluções, processos, atividades de automação e empresas, como processadores, memórias, videotexto, fibra óptica, etc.

Ainda nas palavras de Le Coadic (1996), o autor discorre sobre os problemas que cruzam as fronteiras históricas das disciplinas tradicionais, ficando evidente os recursos utilizados para tal ato, denominando-se interdisciplinaridade, que é regida por meio de uma participação colaborativa entre diversas áreas, gerando interações de modo que ambas as partes adquiram conhecimento, ou seja, que haja um enriquecimento mútuo.

A Ciência da Informação se faz presente nesse âmbito, sendo

> uma dessas novas interdisciplinas, um desses novos campos de conhecimento onde colaboram entre si, principalmente a psicologia, a linguística, a sociologia, a informática, a matemática, a lógica, a estatística, a eletrônica, a economia, o direito, a filosofia, a política e as telecomunicações (LE COADIC, 1996, p.22).

Como seres humanos, somos capazes de aprender sobre o que está à nossa volta de diversas maneiras: observando, lendo, ouvindo e

experimentando, todas estas ações contribuem para a nossa construção de conhecimento individual. Porém, diante dessa avalanche de informações que está presente nas redes de comunicação e, principalmente, acessíveis sem nenhum custo extra, a confiabilidade é um fator que pode perder força em meio a tanto volume informacional. No entanto, quando o conhecimento sobre um determinado fenômeno é obtido seguindo uma metodologia científica, isto é, de acordo com etapas e regras controladas, cria-se um vínculo mais confiável para chegar no resultado, resultando em um conhecimento denominado conhecimento científico ou ciência (KERLINGER, 1979).

Com base nesse cenário, Mueller (2000) relata que a confiabilidade é uma das características mais importantes da ciência, pois consegue distinguir o conhecimento popular do conhecimento científico, construindo, para isso, uma rigorosa metodologia científica para a geração do conhecimento, bem como a divulgação dos resultados para serem julgados por outros cientistas, seus pares.

Hernandes-Cafíadas (1987) relata que, antigamente, a ciência era desenvolvida a efeito por eruditos, onde não havia a intenção de seguir um objetivo definido. Contudo, atualmente, a maioria das pesquisas são desenvolvidas considerando o progresso e a apresentação de uma solução para os objetivos estipulados, que, por sua vez, geram benefícios à sociedade. Dessa forma, as atividades que se tornarem socialmente importantes devem entrar em consenso com a sociedade em questão. A expansão da ciência seguindo esse padrão pode ser utilizada para reformular políticas governamentais e determinar metas sociais, garantindo uma responsabilidade a mais para a ciência, isto é, toda abordagem para o termo deve interagir também com a sociedade.

Para Ziman (1981), a ciência constitui-se de um conjunto de conhecimentos públicos impactados pelas atividades coletivas desempenhadas por pesquisadores, de forma a acrescentar sua contribuição pessoal, corrigida e refinada pela crítica recíproca, em uma ação colaborativa e competitiva com a dos demais contemporâneos.

Bauman, falecido em janeiro de 2017, aos 91 anos, dedicou sua vida a estudar a condição humana, considerado um dos expoentes da chamada “sociedade humanística”, o filósofo, sociólogo, professor e escritor polonês cunhou o termo sociedade líquida para o conceito de pós-modernidade. A

metáfora “líquido” ou da “fluidez” foi escolhida devido ao aspecto do estado das mudanças, se alterando constantemente e não mantendo sua forma por longos períodos. A contemporaneidade assemelha-se pela vulnerabilidade e fluidez, sendo incapaz de manter sua identidade, reforçando a fragilidade e o estado temporário das relações sociais.

Com o avanço da tecnologia, tornou-se possível agir sem sair do lugar, em um ritmo acelerado e constante, de forma a expor a sociedades a riscos que não existiam em outras épocas, como por exemplo a previsão de condições climáticas, a poluição e a diminuição de fontes de energia não-renováveis. Com isso, criam-se relações líquidas, que podem ser rompidas e transformadas em relações solitárias a qualquer momento, enfraquecendo a solidariedade e estimulando a insensibilidade para com os outros indivíduos da sociedade. Ainda nesse contexto, acredita-se que em uma estrutura de laços em rede e não em comunidade, onde cada indivíduo tem o papel de nova conexão com um propósito específico, podendo ser facilmente removido desta estrutura (BAUMAN, 2003).

Corroborando com o pensamento e definições de sociedade líquida de Bauman (2003), Mourin (2000) relata que que o desenvolvimento e a transformação da sociedade se dão por meio da ciência, que, por sua vez, está se desenvolvendo mais rápido e tornando-se cada vez mais perceptível para as toda a comunidade, inclusive a científica. Nesse contexto, Mueller (2000) discorre sobre a continuidade de pesquisas científicas retomadas por outros pesquisadores, teóricos ou aplicados, contribuindo com o avanço da ciência ou com o desenvolvimento de produtos embasados nestas pesquisas.

Mueller (2000) relata que o avanço da tecnologia, especialmente computadores e redes eletrônicas impacta diretamente nos meios de comunicação disponíveis à comunidade científica, tornando-se cada vez mais eficientes, rápidas e abrangentes, de modo a vencer barreiras geográficas, hierárquicas e financeiras, o que vai ao encontro dos pensamentos de Bauman (2003) e Mourin (2000).

Esta seção apresentou os conceitos históricos e abordagens da comunicação científica de modo geral, a próxima subseção apresenta as relações da comunicação científica com a ciência.

3.2 Conceitos e relações da comunicação científica com a ciência

A autoria do termo comunicação científica é dada por John Desmond Bernal, irlandês, físico e historiador da ciência na década de 40 (CARIBÉ, 2015). Em 1939, o autor dedicou um capítulo ao tema em seu livro denominado `A função *Social da Ciência´.* Para Bernal (1939), as atividades associadas à produção, disseminação e uso da informação, desde sua concepção até a informação referente aos resultados alcançados estão relacionadas com o conceito de comunicação científica.

Na literatura especializada, pôde-se observar variações para definir as relações, os processos, a origem e a natureza da comunicação científica. Caribé (2015) contribui significativamente para identificar estas variações por meio de uma técnica denominada análise documental, resultando nos seguintes termos: alfabetização científica, analfabetização científica, compreensão pública da ciência, comunicação científica, comunicação pública da ciência, cultura científica, difusão científica, disseminação científica, divulgação científica, educação científica, jornalismo científico, percepção pública da ciência, popularização da ciência e vulgarização da ciência.

Há séculos, o conhecimento científico prova suas virtudes de verificação e descoberta em relação a outros modos de conhecimento, carregando consigo, de forma singular, o progresso do saber e da construção do conhecimento, sendo parâmetro principal de grandes conquistas inéditas, como a domesticação da energia nuclear e os princípios da engenharia genética. Diante disso, a ciência é elucidativa, isto é, soluciona mistérios, enriquecedora, ou seja, satisfaz necessidades sociais e incentiva o avanço da civilização e, por fim, conquistadora e triunfante (MORIN, 2000).

É válido ressaltar que a ciência se torna ainda mais útil quando suas contribuições alcançam não somente o público interno, isto é, aqueles que atuam diretamente nos trilhos científicos, mas sim ao público externo, de modo que as pessoas consigam associar o valor da ciência para a resolução de problemas e garantia de segurança nas atividades presentes na vida do ser humano. Nesse contexto, Caribé (2015) aborda a institucionalização da comunicação científica na comunidade científica de maneira fluida, entretanto, essa fluidez carece quando se refere ao público externo à comunidade

científica, ponto que pode estar relacionado com as leis e regras internas que regem esse grupo social.

Nesse contexto, Anna (2019, p.2) relata que

> [...] a comunicação científica representa um dos pilares básicos e mais importante para as ciências, pois ela garante, além da comprovação das descobertas, a sua aceitação por pares e, como consequência, sua divulgação e legitimidade em meio aos grupos que comungam de ideias semelhantes, por decorrência, essas descobertas são compartilhadas, divulgadas e utilizadas em benefício da sociedade. (ANNA, 2019, p.2.)

Silva (2001), ressalta que, do ponto de vista cognitivo, um novo somente adquire seu valor quando ele é difundido dentro da comunidade, desta forma, contribuindo com o avanço científico. Do ponto de vista social, a publicação de novos descobrimentos é uma etapa primordial do processo investigativo, o que permite ao cientista o reconhecimento de sua própria obra.

Observa-se no âmbito científico que os pesquisadores recebem informações essenciais para o desenvolvimento de seus trabalhos ou atividades cotidianas. Nesse contexto, Bernal (1939) aborda a importância de cientistas e leigos receberem estas informações, construindo um conceito sobre comunicação científica, de modo a segregar em duas frentes: o aspecto interno em que a comunicação se sobressai diante da comunidade científica e o aspecto externo à comunidade científica, denominada pelo autor como educação científica e popularização da ciência, ambas abordagens subsidiadas pela ampla abrangência em que se dá o termo comunicação científica.

No que tange à comunidade científica, existem diversas definições na literatura para definir o termo. Le Coadic (1996) define como:

> [...] redes de organizações e relações sociais formais e informais que desempenham várias funções. Uma das funções dominantes é a de comunicação. O papel da comunicação consiste em assegurar o intercâmbio de informações sobre os trabalhos em andamento, colocando os cientistas em contato entre si (LE COADIC, 1996, p.33.)

Caribé (2015) aborda a comunidade científica como um grupo responsável pela produção de conhecimento científico, constituído em um grupo social bem definido, com regras e características individuais.

Para Schwartzman (2001), a comunidade científica pode ser entendida como um grupo de indivíduos que compartilham valores e atitudes científicas, comunicando entre si por meio de suas instituições científicas. Sua formação é dada por indivíduos que possuem em comum habilitações, conhecimentos e premissas tácitas sobre uma determinada área do conhecimento, onde cada indivíduo conhece sua área de atuação, ressaltando que que ninguém possui um conhecimento exaustivo e sistemático de todo o campo. Aborda-se ainda, sobre um sistema de autoridade que é responsável por defender os critérios de probidade, plausibilidade e aceitabilidade dos resultados, tornando-se parte integral e fundamental do funcionamento de um método científico, apesar de não constituírem um traço explícito e obrigatório do método em questão.

Na sociedade moderna, Fourez (1995) define a comunidade científica como um grupo social relativamente bem definido, onde sua estrutura dá-se por si mesmo, de modo que os indivíduos fazem seu próprio reconhecimento de um mesmo corpo. Nesse contexto, pode-se observar uma aproximação de outros grupos sociais do considerando a antropologia e a sociologia: os sapateiros, os alquimistas ou os feiticeiros. Nesses casos, há a autodefinição dos indivíduos pertencentes a esses grupos, gerando um reconhecimento e coerência mútua entre os envolvidos.

Fourez (1995) ressalta ainda que a comunidade científica se difere dos alquimistas considerando sua aceitação por toda a sociedade ou boa parte dela. Os indivíduos que se enquadram como cientistas, possuem o poder do conhecimento, sendo passível de distribuição. Diante disso, a comunidade científica usufrui não apenas de reconhecimento interno, mas de conhecimento externo, apesar da carência de divulgação científica apontada por Caribé (2015).

Observa-se que a exposição das pesquisas científicas para o público leigo, ou seja, todos os indivíduos externos à comunidade científica é um ato bem-visto por muitos pesquisadores, contribuindo com a evolução da sociedade e gerando maior confiabilidade nas informações que estão sendo passadas para toda a sociedade, já que todos os trabalhos científicos requerem aprovação por outros membros da área. Nesse contexto, Fourez

(1995) relata que a sociedade se difere da comunidade científica, pois há uma complexidade na identificação dos responsáveis quando surge alguma afirmação originada a partir da comunidade científica ou sociedade. Contudo, quando se faz referência à comunidade científica, é sabido que os envolvidos devem ser todos aqueles que detém o poder do conhecimento, do mesmo modo quando são abordados os interesses, culturas e regras da sociedade.

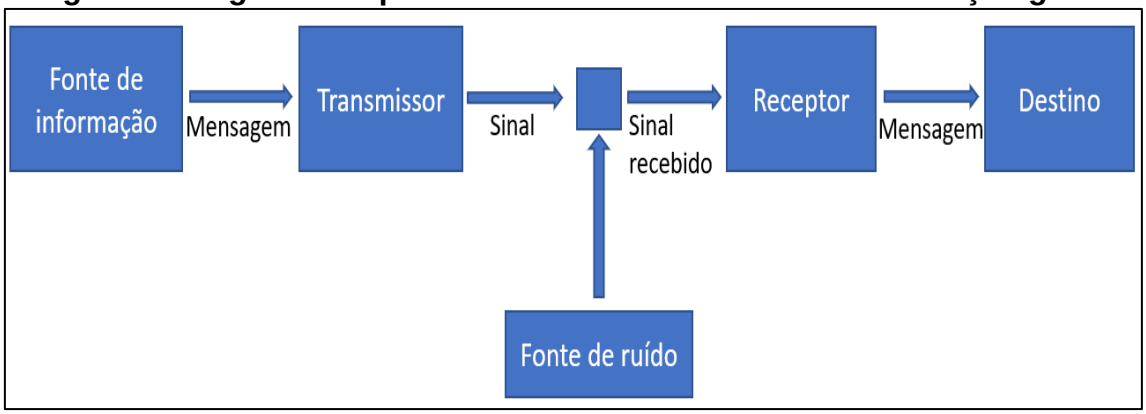
Considerando ainda o conceito de comunidade científica, Fourez (1995) reforça a ideia de que não se pode definir o termo como apenas um grupo capaz de lidar com certos tipos de conhecimentos. Um grupo científico poderá ser solicitado com certa frequência para comentar sobre um determinado assunto, de modo que suas opiniões sejam consideradas especialistas no contexto, tornando-os detentores de conhecimento que lhes permitem opinar sobre a sociedade.

Miranda (1996) aborda as atividades relacionadas com a produção de conhecimento científico de duas maneiras, primeiramente aquelas que auxiliam no processo de viabilizar um produto enquanto veículo, ou seja, que forneça suporte físico de comunicação do conhecimento, posteriormente, atividades que têm ligação com o produto e nos fornecem parâmetros para propor inferências acerca da comunicação entre os envolvidos (cientistas) de uma comunidade científica, de modo que a interação não se infere pela interação, mas pelo produto final desta – os artigos científicos. Nesse contexto, como atividades associadas à disseminação, entende-se aquelas que conferem tanto ao produto, quanto aos produtores a projeção necessária à sua visibilidade no meio social em que esses indivíduos estão inseridos.

Caribé (2015) ressalta que a comunicação científica pode ser entendida como um “processo de comunicação clássico”, gerando conteúdo informacional que, por sua vez, é decorrente dos processos intrínsecos ciência, desenvolvidos por cientistas, pesquisadores, acadêmicos e outros profissionais envolvidos em diferentes áreas do conhecimento. Shannon e Weaver (1948) reforçam esta abordagem, onde têm-se o emissor, a mensagem, o canal e o receptor, conforme se apresenta na figura 3.

De acordo com Fourez (1995), os indivíduos necessitam de conhecimento para que eles possam tomar suas próprias decisões, tal necessidade deve ser amparada pela comunicação científica, de modo que o conhecimento científico chegue até o público leigo.

Figura 3: Diagrama esquemático de um sistema de comunicação geral



Fonte: Shannon e Weaver (1948, p.2 – tradução nossa).

A figura 3 demonstra a arquitetura de um sistema de comunicação que consiste em 5 partes segundo Shannon e Weaver (1948) – tradução nossa.

1. Uma fonte de informação que produz uma mensagem ou sequência de mensagens a serem enviadas ao terminal de recepção.
2. Um transmissor, responsável por operar na mensagem de alguma forma para produzir um sinal adequado que posteriormente será transmitido pelo canal.
3. O canal, que é apenas o meio utilizado para transmitir o sinal do transmissor ao receptor. Pode ser um par de fios, um cabo coaxial, uma banda de frequências de rádio, um feixe de luz, etc.
4. O receptor, que tem o papel de normalizar a operação inversa à operação feita pelo transmissor, reconstruindo a mensagem do sinal.
5. O destino, sendo a pessoa ou objeto que receberá a mensagem.

Observa-se que tais processos podem ser assimilados com a realidade da produção científica atualmente, por meio de uma analogia periférica, onde a **fonte** de informação são os cientistas ou responsáveis pela publicação de conteúdo científico, os **transmissores** podem ser vistos como as bancas de avaliação destes materiais, onde é requerido uma série de critérios para atender aos padrões da revista ou repositório de conteúdo científico. O **canal** é a internet, juntamente com toda sua infraestrutura de segurança e usabilidade. O **receptor**, pode ser entendido como os profissionais e cientistas que submeteram o material para publicação, responsáveis por fazer a adequação

do material submetido para adequar-se aos padrões do destinatário. Por fim, o **destino** são os usuários que consomem a mensagem, ou seja, todos aqueles que irão baixar os materiais científicos de alguma fonte especializada.

Hernández Cañadas (1987) discorre sobre a forma de transmissão das informações, dando origem a dois canais denominados canais informais ou canais formais de comunicação. No contexto da estrutura da comunicação, os receptores podem variar, podendo ser considerados tanto cientistas, como colegas de trabalho ou pares, como pessoas externas à comunidade científica, ou seja, o público em geral (leigo). Ainda nesse contexto, corroborando com a estrutura do sistema de comunicação apresentado anteriormente na Figura 3, relata-se que o ato de comunicar “pode estar formado ou considerar os seguintes fatores: quem: fonte; o que: mensagem; de que forma: canal; para quem: receptor” (HERNÁNDEZ CAÑADAS, 1987).

Seguindo o contexto de encontrar na literatura abordagens relacionadas à comunicação científica, no que tange à Ciência da Informação, há profissionais que corroboram e referenciam-se com base no texto de William D. Garvey, publicado em 1979. Garvey (1979) aborda o termo de forma mais aberta, expondo seus conceitos e aplicações para os demais públicos. Utilizando-se como base o trabalho de Bernal (1939), o termo foi definido como um conjunto de atividades que se relaciona com a produção, disseminação e uso da informação desde a concepção da ideia de um cientista até a confirmação de que a pesquisa foi aceita como parte constituinte do conhecimento científico.

O sistema de comunicação científica envolve diferentes aspectos da ciência, entre eles, a comunicação entre pesquisadores, com o intuito de anteceder teorias acerca de um tema específico pode ser chamada de comunicação formal ou comunicação informal. A comunicação formal, como o próprio nome já diz, utiliza canais formais, utilizados em casos de publicações com divulgação mais ampla, como periódicos e livros. A comunicação informal, faz uso de canais informais e pode incluir caráter mais pessoal ou que se referem a outros trabalhos em andamento, como comunicação de pesquisas em andamento e atividades semelhantes (MUELLER, 2000).

Com base nos conceitos anteriores, o conhecimento científico pode ser gerado por meio de canais formais e informais, podendo ser resultado de apenas uma pesquisa ou continuação de um trabalho de maior amplitude, fato

que é comum na comunidade científica, por isso a importância de uma boa metodologia científica, para que aqueles que tiverem interesse em continuar um determinado trabalho, tenham uma base confiável para garantir que seus resultados sejam corretos.

Mueller (2000) relata que o período em que uma inovação científica permanece inédita tem sido cada vez menor, ou seja, o tempo em que a publicação inicial divulga seus resultados e publicações posteriores que obtiveram melhorias em relação a estes resultados apresentam suas conclusões está ficando cada vez mais curto. Isso pode acarretar problemas para todos os interessados no tema, já que a publicação ainda recente pode cair em desuso, criando uma complexidade para os cientistas se manterem atualizados e informados e para os centros de informação, do ponto de vista financeiro, uma vez que suas bases de dados crescem exponencialmente.

Nesse contexto, Targino (2000) relata que a ciência busca desvendar e compreender a natureza e seus fenômenos por meio de métodos sistemáticos e confiáveis. Contudo, a dinamicidade em que se dá os atributos intrínsecos à ciência, como os periódicos científicos por exemplo, seus resultados são sempre temporários, levando a ciência a adentrar-se em um processo investigativo, tornando a ciência uma instituição social e dinâmica, de forma a exercer atividades que contribuam com a construção do conhecimento.

Observa-se que a ciência contribui com o avanço da sociedade utilizando o avanço da tecnologia para alavancar ainda mais sua conduta perante a comunidade, orientando a evolução humana e distinguindo resultados verdadeiros dos que não possuem uma validação científica mais aprimorada. Muitas tarefas são executadas mediante consulta em materiais científicos, os quais possuem um grau de confiabilidade obtido por meio de revisão crítica do conteúdo abordado, garantindo a usabilidade e consulta para outros indivíduos, sejam eles, cientistas ou o público leigo. No entanto, Targino (2000) ressalta que a verdade e a certeza absoluta inexistem, garantindo aos cientistas somente o papel de identificar a verdade, mas não a deter.

Diante de tais abordagens, observa-se que alguns autores discorrem sobre o fato de um autor ser referenciado em um determinado trabalho não garante que ele seja o detentor da verdade, isto é, o indivíduo que possui o maior e mais correto arcabouço teórico sobre o tema. Por outro lado, referenciar um determinado autor muitas vezes em trabalhos direcionados para

a mesma área pode nos levar a crer que tal pessoa realmente detém maiores conhecimentos para discorrer sobre o assunto.

Ainda sobre os conceitos voltados para canais formais e informais, Silveira (2000) relata que, por meio destes canais, a comunicação científica, bem como o conhecimento científico pode ser produzido de duas maneiras: (i) a comunicação científica e (ii) a comunicação pública da ciência, onde ambas as possibilidades são constituídas por canais distintos, porém, o que os distingue é uma linha tênue e, muitas vezes, imbricada.

Para Silveira (2000), os canais de comunicação (formal e informal) “são subsistemas da estrutura do sistema de comunicação da ciência e caracterizam-se por diferenças dentro do processo de produção, disseminação e uso da informação” (SILVEIRA, 2000). Hernández Cañadas (1987) relata que em cada um destes canais estão contidos processos diferentes de produção, disseminação e uso da informação.

Esta seção apresentou alguns conceitos sobre a comunicação científica, com o intuito de subsidiar conceitualmente o conteúdo que abrange a temática da pesquisa. A próxima seção tratará sobre o contexto da citação.

4. CITAÇÃO: CONCEITOS E ABORDAGENS

Esta seção aborda os conceitos, definições, aplicabilidade, teorias e contextos socioculturais das citações, buscando apresentar suas representações epistemológicas e opiniões de diferentes autores na literatura sobre o uso de citações em trabalhos científicos.

4.1 Definição e aplicação das citações

Os estudos de citação fazem parte de um ecossistema científico que envolve diferentes áreas do conhecimento, sendo realizados pelos pesquisadores como um fator essencial para o campo da ciência. Com a multidisciplinaridade sendo amplamente utilizada em muitas pesquisas, os atos de citação e referência podem contribuir de maneira significativa com as pesquisas em diferentes áreas do conhecimento.

De acordo com Silveira e Caregnato (2017, p.152), os estudos de citação podem ser compreendidos como

> investigações teóricas, metodológicas e aplicadas direcionadas às práticas de citação, suas relações e de seus componentes (registros citados e referenciados), em contribuições científicas certificadas por meio de técnicas quantitativas e qualitativas, com o propósito de evidenciar, observar, entender e analisar a dinâmica da produção, comunicação e uso do conhecimento. (SILVEIRA e CAREGNATO, 2017, p.152).

No Brasil, os estudos de citação se deram por meio de iniciativas conduzidas por pesquisadores da Ciência da Informação há pouco mais de quarenta anos, baseando-se nas leis bibliométricas. O aumento da produção nesse contexto se deu principalmente na última década, porém, em períodos específicos, pode-se notar uma certa pausa. Para tanto, os trabalhos científicos amadureceram no que tange à teoria e metodologia, amparando-se em tecnologias de informação que potencializaram o desenvolvimento de técnicas específicas e a visualização de resultados. Tal crescimento poderia ser consequência do aumento do número de trabalhos, a sofisticação de *softwares* e das bases de dados e o aumento de programas de pós-graduação no país, porém, tais artefatos não indicam um avanço epistemológico significativo para os estudos de citação, considerando as últimas três décadas (Silveira e Caregnato, 2017).

De acordo com a I4OC (2022), as citações são os elos que unem o conhecimento científico e cultural. Além disso, são dados primários que fornecem tanto a proveniência quanto a explicação sobre como temos conhecimento dos fatos.

A palavra citação é amplamente empregada para descrever o ato de remeter um artigo para outro, onde tem-se o artigo citante, que possui a referência, e o artigo citado, que foi mencionado nas referências. Quase todos os artigos científicos trazem consigo referências de publicações afins, com o intuito de justificar argumentos, criticar trabalhos anteriores, etc. Essas referências podem ser vistas como um mecanismo de rede que integra de modo geral a literatura científica (MEADOWS, 1999).

De acordo com a Associação Brasileira de Normas Técnicas (2003, p. 2), ‘citação’ define-se por “menção de uma informação extraída de outra fonte”.

Alvarenga (1998, p.6) relata que o ato de citar é permeado por

> todo um espectro de implicações psicológicas, sociológicas, políticas e históricas, assim como influências de outras naturezas, tais como o narcisismo (autocitações), influências entre autores e instituições, adesão a paradigmas vigentes. Nas práticas discursivas, o hábito de citar ou fazer referência a um trabalho anteriormente escrito pode ser considerado parte constitutiva do processo de enunciação ocorrida em campos específicos dos saberes (ALVARENGA, 1998, p.6).

Hoffnagel (2009) discorre sobre as funções atribuídas à citação, como por exemplo, permitir que o escritor se apoie em outros autores para sustentar seus argumentos. Ademais, a reputação do autor citado pode ser reivindicada como suporte para o trabalho do escritor citante. Ressalta-se também que reconhecer embasar seus argumentos em outros estudos pode passar uma visão fiel e hábil sobre sua imagem, onde buscou-se identificar e analisar estudos de outros pesquisadores, podendo gerar maior credibilidade.

Nessa perspectiva, Carvalho (1975) cita algumas funções das citações bibliográficas no contexto da Comunicação Científica:

- Contribuir para o desenvolvimento da ciência;
- Fornecer o reconhecimento de um cientista por seus colegas;
- Determinar os direitos de propriedade e prioridade da contribuição científica de um autor;
- Compor importantes fontes de informação;
- Auxiliar a julgar as formas de coleta de informação;
- Apresentar a literatura que é indispensável para o trabalho dos cientistas.

Erikson e Erlandson (2014) sugeriram quatro categorias de motivos para integrar o ato de citar: Argumentação, Alinhamento Social, Alinhamento Mercantil e Dados. As três primeiras categorias são mais comumente encontradas em grande parte dos artigos científicos. A última, porém, limita-se a trabalhos que objetivam analisar outros trabalhos, como artigos de revisão ou estudos meta-analíticos.

Vieira (2010) ressalta que a utilização de um trabalho externo, o uso de trechos inteiros ou somente ideias contidas no trabalho pesquisado requer que ele seja citado, com o intuito de evitar plágio. Diante disso, ressalta-se que em situações em que se considera a ideia de outro autor, mesmo utilizando outras palavras, exige a necessidade da citação. Para o autor, a principal finalidade

das citações é consolidar um argumento ou ponto de vista do próprio pesquisador, baseando-se em ideias de alguém mais especializado no tema.

A primeira categoria, denominada argumentação, geralmente contém as funções mais tradicionais relacionadas à citação, onde o ato de citar é utilizado em uma linha de argumentação para apoiar um determinado ponto de vista. Esta categoria divide-se em cinco subcategorias de argumentação: delimitação, suporte ativo, crítica ativa, suporte passivo e leitura adicional. A segunda categoria, chamada Alinhamento Social, se dá pela identidade ou autoconceito do autor. Neste caso, divide-se em três subcategorias: tradição científica, autoimagem científica e compensação de esforço. As subcategorias referem-se às formas com que o autor citante se apresenta durante a escrita do texto, além de proporcionar maior segurança de um campo bem definido. A terceira categoria, denominada Alinhamento Mercantil, origina-se a partir da obtenção de créditos de vários tipos por parte dos autores. Aplica-se também a combinação de normas mertonianas com discussões construcionistas de posições de influência. Para isso, têm-se três subcategorias: Crédito, Credenciais Próprias, Material de Troca, Autopromoção e Penhor. Por fim, a quarta categoria, tida como Dados, diferencia-se das três anteriores, pois, neste caso, a literatura utilizada transforma-se em dados para ser trabalhada pelo autor citante. Para tanto, subdivide-se em três categorias: Revisão, Meta-análise e Estudos de texto (ERIKSON; ERLANDSON, 2014).

Erikson e Erlandson (2014) ressaltam ainda que o conjunto de categorias desenvolvido não tem o objetivo de ser a referência final para o tema. Na medida em que a prática social da construção do conhecimento evolui, bem como os padrões de publicação e os modelos de avaliação tenham mudanças, podem surgir novas categorias ou métodos de se categorizar os motivos de realizar uma citação, abordando ainda a instabilidade como um ponto que não deve ser considerado como fraqueza no modelo estabelecido, mas sim enriquecedor, uma vez que, aumenta ainda mais a complexidade da citação.

Há a possibilidade de se mesclar, combinar ou dividir de outras maneiras, pois, a citação possibilita que as relações semânticas entre um artigo e os documentos citados sejam identificadas. Para isso, considera-se que os novos artigos que citam os mesmos documentos publicados anteriormente possuam geral relações semânticas entre si.

Em outra linha de raciocínio, pode-se dizer também que a citação e a referência a outros conteúdos se integram com relevância do processo de produção científica, seguindo as orientações requisitadas que tendem a vincular o conhecimento construído aos conceitos, paradigmas ou fenômenos vigentes em uma determinada área do conhecimento, atribuindo-lhe uma possível visão de autenticidade imposta pela obediência aos imperativos institucionais dessa área do conhecimento (ALVARENGA, 1998).

O comportamento dos pesquisadores em relação à citação é um tema abordado por cientistas para entender como as obras são construídas em cada área do conhecimento e observar suas variações. Para Rousseau (1988), essas variações tornam os estudos de citação fundamentais para a compreensão da comunidade científica.

No que se refere à análise de citações, com base na premissa de que os pesquisadores submetem seus trabalhos considerando obras anteriores, provando isso citando obras precedentes em formato ordenado e padronizado de referências. Diante disso, torna-se evidente o comportamento dos cientistas a partir do estudo dessas citações (MOREL; MOREL, 1977).

Dessa forma, considerando a importância das referências bibliográficas em um trabalho científico, a análise dessas referências, que são denominadas no campo da bibliometria como Análise de Citação, vêm sendo empregadas como um importante instrumento metodológico de mapeamento da produção intelectual de diversas áreas do conhecimento. Com isso, a utilização da Análise de Citação serve de parâmetro para diferentes finalidades, como a indicação de tendências de temáticas de pesquisa, indicadores de citação e o mapeamento de áreas do conhecimento mais citadas (MORAES, FURTADO e TOMAÉL, 2015).

Diante da ampla quantidade de referências sobre o conceito do termo citação, há também a necessidade de apresentar os tipos de citação existentes. Diante disso, a próxima subseção aborda as diferentes formas de citar um trabalho.

4.2 Tipos de citação

Existem algumas formas de citar um material dentro do texto, entre elas, Vieira (2010) relata que a citação deve respaldar ou acrescentar argumentos que solidifiquem a exposição prévia ou posterior apresentada pelo pesquisador, distinguindo-se em duas formas: direta e indireta.

De acordo com Henriques e Medeiros (2017, p. 190) afirmam que a citação

> [...] pode ser direta ou indireta. A direta constitui-se na transcrição de texto de outro autor. A indireta consiste numa paráfrase do texto do autor consultado, ou seja, texto baseado na obra do autor consultado. A citação pode ser ainda de outra citação. Nesse caso, embora não se tenha tido acesso à obra de um autor, transcrevem-se suas ideias, consultando não o livro original, mas uma citação em texto de terceiros. Portanto, trata-se de citação direta ou indireta de um texto em que não se teve acesso ao original. Esse tipo de citação é representado pela expressão latina apud. Recomenda-se que seu uso se restrinja a textos raros (HENRIQUES E MEDEIROS, 2017, p.190).

Vieira (2010), possui uma mesma linha de raciocínio, onde as citações diretas se fazem necessárias em situações em que se transcreve, literalmente, o conteúdo para o trabalho, de modo que se referencia sua origem. Para o autor, esse tipo de citação deve ser usado para provar que um determinado autor afirmou algo de maneira inequívoca, a fim de evitar más interpretações das palavras do autor. Entretanto, o uso de citações diretas pode requerer do pesquisador explicações que a coloquem dentro de um contexto mais geral do que está sendo citado, de modo a prevenir que sua utilização não seja para comprovar proposições distintas daquilo que está sendo citado, tendo por consequência, a prática de fraude.

Henriques e Medeiros (2017) discorrem sobre os tipos de citações existentes, entre elas, têm-se:

- **Citação direta**: trata-se de uma citação em primeira mão. Neste caso, se a citação tiver até três linhas, deve ser apresentada entre aspas duplas (“), no mesmo parágrafo em que o pesquisador vem expondo seu raciocínio. Caso a citação tenha mais de três linhas, deve-se isolar a citação, de forma a destacá-la por meio de normas técnicas do parágrafo atual;

- **Citação indireta**: trata-se de uma citação em segunda mão, tira-se de um autor, por intermédio de outro. Para o autor, justifica-se este uso quando a obra citada por terceiros é inacessível ou há uma dificuldade em encontrar sua localização, influenciada por questões como autores antigos, estrangeiros ou edições esgotadas;
- **Citação literal**: trata-se de uma citação que segue exatamente a forma com que o autor escreveu, isto é, mesmo se tiver erros ortográficos ou gramaticais. No caso de erros, emprega-se a palavra *sic,* que pode ser utilizado também para sinalizar contradições, inadequações, estranhezas ou ironia;
- **Citação parafraseada**: neste caso, trata-se de uma transcrição do pensamento do autor por meio de vocabulário e estilo próprio do autor que está citando. Este tipo de citação é considerado como citação indireta;
- **Citação condensada**: trata-se de uma citação indireta. Consiste no resumo ou na síntese de um material consultado sem que altere a ideia original do autor citado, de modo que não apresente juízos de valor ou comentários de ordem pessoal.

Ainda em relação às citações indiretas, sua forma de apresentação se dá por meio de paráfrases, onde se tem mais elegância ao utilizar esta abordagem, pois condensam mais o trabalho em relação ao simples fato de fazer uso de citação direta, onde cita-se literalmente um autor. Outro ponto abordado pelo autor, é que o uso de citações indiretas demonstra que o pesquisador interagiu com mais de uma pesquisa sobre o mesmo tema e conseguiu interpretar e separar argumentos que contribuíssem com suas ideias em diferentes fontes de pesquisa, além demonstrar sua capacidade de fazer analogias e a síntese de ideias, consequentemente, amadurecendo-o como pesquisador. Ressalta-se também que as citações indiretas ocupam um espaço físico menor em comparação às citações diretas (VIEIRA, 2010).

Esta seção buscou contextualizar sobre alguns tipos de citação presentes na literatura. É sabido que existem outras abordagens sobre áreas

distintas, como direito, por exemplo. No entanto, buscou-se apresentar as mais utilizadas em artigos científicos.

4.3 Perspectivas sobre o uso de citações

Pode-se observar que alguns trabalhos possuem mais referências e outros contam com um número mais limitado, o ponto é que delimitar um número ou padrão para referências bibliográficas de uma obra científica é uma tarefa complexa e rodeada de questionamentos, por exemplo: (1) quanto mais referências um trabalho tem, mais confiável ele é? (2) o fato de um autor ser muito citado em diferentes trabalhos, sejam eles da mesma área ou não, é parâmetro para dizer que é a pessoa mais apta para abordar um determinado assunto? (3) ao contrário do item anterior, um autor pouco ou nada citado pode ser considerado inapto ou pouco confiável para o contexto? (4) trabalhos científicos que possuem referências mais antigas são os mais apropriados? (5) trabalhos que possuem referências mais atuais são totalmente confiáveis? Existe um padrão ou orientação para esse tipo de situação, e, caso exista, qual o embasamento científico para afirmar tal coisa? (6) motivos pelos quais autores são citados apenas em partes específicas, como início, meio ou fim. Existe algum ponto positivo ou negativo em ser citado somente em uma parte do texto?

Todas essas questões possuem uma natureza complexa e requer muita investigação científica para chegar a uma possível resposta, e, talvez, não definitiva. Braga (1974) relata que um campo científico citar trabalhos mais antigos do que a literatura atual indica um tipo de “metabolismo humanístico”, sendo necessário digerir tudo o que já foi publicado, amadurecer o conhecimento adquirido para possibilitar a produção de novos textos, que possivelmente abordarão os mesmos tipos de problema.

De acordo com I4OC (2022), as citações permitem atribuir e creditar contribuições científicas, o que possibilita avaliar pesquisas e seus possíveis impactos. Em outras palavras, as citações são o veículo mais importante para a descoberta, disseminação e avaliação de todo o conhecimento acadêmico.

Nesse contexto, Carvalho (1975, p. 119) relata que:

> Não se pode esperar que todos os autores sejam cuidadosos, objetivos e conscientes no momento de mencionar suas fontes

> de consulta. Alguns pecam por excesso, outros por omissão. Vários fatores podem influenciar os autores na escolha das citações de seus trabalhos. Há autores de renome num campo, que são citados para realçar o trabalho de quem os cita. Há autores que são escolhidos para que a responsabilidade em assuntos controvertidos seja dividida. Há citações que indicam o apreço a colegas, hostilidade a concorrentes ou obediência à política editorial. A possibilidade de um documento ser citado dependerá também da acessibilidade, da procedência (país onde foi originalmente publicado), da língua, do tipo de material bibliográfico e da data de publicação (CARVALHO, 1975, p.119).

Hoffnagel (2009) relata que a prática de citação de trabalhos anteriores no desenvolvimento de novos trabalhos pode ser uma atividade obrigatória, porém, tal atividade não é simples de ser realizada, pois dependem de escolhas retóricas complicadas. O cientista precisa decidir o momento exato para fazer uma citação, bem como decidir quem será o autor citado, além de aplicar o tipo de citação (direta ou parafraseada).

Braga (1974) discorre sobre o surgimento dos índices de citações, especificamente *Science Citation Index*, onde foi possível elaborar estudos aprofundados em diferentes tipos de literatura com base em métodos de contagem de citações, onde é possível determinar se um determinado campo do conhecimento (ou documento) comporta-se como “Ciência” ou “Não Ciência”. Entretanto, tais ferramentas ainda não permitem que se estabeleça limites radicais para uma obra científica, como demarcar documentos com doze citações como mais eruditos em comparação aos que contém apenas dez.

Em relação a definir o que é ciência do que não é, considerando o número de citações como parâmetro, Braga (1974, p.174), relata que:

> O número de citações, isoladamente, não é um parâmetro bastante preciso para distinguir Ciência de Não-Ciência. É necessário medir a natureza do sistema de citações; é preciso conhecer, além do crescimento e da densidade procriativa do sistema, seu metabolismo ou eugenia (BRAGA, 1974, p.164).

Carvalho (1975) corrobora com Braga (1974), ao dizer que grande parte do aumento das pesquisas sobre citações bibliográficas se deu mediante lançamento do *Science Citation Index,* em 1963, pelo *Institute for Scientific Information,* o qual possibilitou um acesso metódico à literatura científica por meio de uma sustentação contínua. Tal fato resultou em um aumento

considerável em pesquisas que visam trabalhar problemas voltados para a comunicação, o que consequentemente aumentou a relevância das citações bibliográficas no domínio da ciência, especificamente nos campos da recuperação da informação, pesquisas de natureza histórica e sociológica.

Uma das questões levantadas por Braga (1974), foi o motivo de documentos científicos possuírem, em média, de dez a vinte e duas citações. Para a autora, não há uma resposta precisa para essa questão. No entanto, pode-se observar que, considerando o índice atual de crescimento da literatura, isto é, 7% ao ano, a quantidade de citações sobre a literatura antiga é equivalente a ordem de magnitude que o conjunto total da literatura antiga. Com isso, um documento antigo ocasiona, em média, uma citação anual, fator que está se transformando lentamente, porém, de maneira persistente, em todos os campos do conhecimento.

O uso de indicadores bibliométricos para estudar as atividades de pesquisa científica se baseia na premissa de que as publicações científicas são um demonstrativo essencial de sua própria presença e qualidade (SILVA, 2001).

Uma das regras da cientometria, de acordo com Barata (2015), é que os indicadores de produção científica acabam modelando comportamentos, como o crescimento no número de publicações, independente da qualidade e direcionamento do material publicado para cenários internacionais, o que, certamente, contribui de forma deletéria ao sistema.

Diante das diferentes opiniões e contribuições dos autores para esse contexto, não se pode afirmar que um indivíduo possui pleno conhecimento e é a pessoa mais adequada para abordar um determinado assunto somente considerando os trabalhos que o citaram. Na opinião de Targino (2000), a verdade e a certeza inexistem, garantindo aos cientistas somente o papel de identificar a verdade, mas não a deter.

Para Silveira e Caregnato (2017), as discussões epistemológicas em torno da ciência concentram-se em delimitar, com clareza e precisão, os elementos que demarcam seus escopos de atuação. Para tanto, muitos destes estudos estão direcionados em compreender e configurar a existência de um objeto científico que será analisado por métodos adequados a questionamentos e hipóteses, conforme as diferentes concepções de ciência.

Nesse contexto, Lalande (1999) considera que a identificação, a compreensão e a construção de hipóteses para um determinado problema são atividades que requerem a delimitação do objeto de estudo. Para tanto, a interação com o objeto é definida baseando-se nas características dos problemas formulados e de suas possíveis resoluções, considerando as variações de observação e análise do objeto em questão.

Bunge (1980) ressalta que a interação com o objeto vai além da contribuição para a definição do método, podendo contribuir também para a condução das análises e verificação de hipóteses, pois, sua representação material da investigação científica pela ciência permite uma constante conexão, acrescentando positivamente na compreensão sobre os fenômenos científicos.

Leydesdorff e Amsterdamska (1990) relatam que os objetos de estudo de citação podem ser vistos na relação entre autores citantes e citados e dos textos citantes e citados, possibilitando caracterizar e visualizar as práticas discursivas que compõem o ato de citação.

Silveira e Santos (2021) reforçam esta ideia, de modo que o ato de citar e referenciar fornece elementos e subsídios para análises e avaliações de práticas de citação dos pesquisadores nas dimensões produtivas e discursivas.

A dinâmica de planejamento, a execução e a apresentação sobre os estudos de citação estão marcadas por estabelecer teorias, metodologias e aplicações que auxiliam a compreensão de realidades científicas. As citações e referências possibilitam que estas atividades forneçam indicadores que revelam a multiplicidade das práticas objetivas e subjetivas, que permeiam o universo científico, por meio de demarcações espaciais, temporais, temáticas, comportamentais, entre outras (SILVEIRA E CAREGNATO, 2017).

A construção desses indicadores é impactada por uma nova área do conhecimento denominada cientometria, consolidada e originada no início do século XX, dedica-se a investigar a ciência e os cientistas, tomando como base os produtos de suas próprias autorias, como patentes, livros e artigos, possibilitando a condução de análises estatísticas para investigar perfis e tendências da própria ciência (LETA, 2011).

De acordo com Price (1969), a cientometria pode ser definida como o estudo quantitativo da atividade científica.

Para Silva (2001, p.2), a cientometria pode ser definida como:

> [..] o estudo da mensuração do progresso científico e tecnológico e que consiste na avaliação quantitativa e na análise das inter-comparações da atividade, produtividade e progresso científico (SILVA, 2001, p.2).

Silveira e Caregnato (2017) relatam que a natureza, as dimensões, a repercussão social e as relações epistemológicas dos estudos de citação reforçam o potencial dos elementos de produção, comunicação e uso do conhecimento que compõem o conjunto de práticas voltadas para o ato científico.

Esta seção buscou apresentar um aporte teórico sobre as definições e aplicação das citações no contexto científico. A próxima seção abordará algumas teorias da citação e o contexto sociocultural que permeia as citações no âmbito da ciência.

4.4 Teorias da citação e o contexto sociocultural

Na literatura, os estudos apontam para duas teorias em relação aos estudos de citação, a primeira, denominada teoria normativa, amparada na ideologia de que a ciência é um campo governado por práticas de citação decorrentes de dívidas intelectuais, recompensas e sanções internas, baseando-se nos preceitos mertonianos e independentes, ou seja, livre de interferências sociais e culturais externas (LEYDESDORFF, 1998; NICOLAISEN, 2007). Ainda em relação à teoria normativa, Merton (1973) relata que nesse caso, há uma troca de crédito na comunidade científica, onde as citações são visualizadas como moedas simbólicas para pagar dívidas intelectuais.

A segunda teoria, chamada de teoria construtivista, possui um foco mais direcionado para identificar os motivos que subsidiam as origens das citações realizadas pelos cientistas, com o propósito de tentar estabelecer a trajetória percorrida para a construção do conhecimento. De acordo com Gilbert (1976), a teoria construtivista pode ser entendida como uma abordagem construtivista social, de modo que não se pode compreender as citações apenas pelo domínio da função intelectual.

Alvarenga (1998) discorre sobre a importância de se desenvolver uma pesquisa bibliometria a partir de literatura caracterizada como formadora de um campo de conhecimento. Para a autora, é importante que o pesquisador tenha,

antes de tudo, a consciência de que a construção do conhecimento deve ser vista como um processo que desconsidera interesses pessoais e subjetividade, devido aos imperativos que governam a conduta social implícita na construção não somente da ciência, mas do conhecimento em geral.

Para Cozzens (1989), ambas abordagens se unem na compreensão de que as citações integram um ponto de encontro entre o sistema retórico de criação de influência e poder e um sistema cuja função se dá pela troca de crédito por trabalho desenvolvido. Nesse contexto, Bornmann e Daniel (2008) ressaltam que mesmo com as duas correntes teóricas contemplando os fenômenos relacionados às citações e referências, nenhuma delas é capaz de solucionar de maneira efetiva e convincente os problemas pelos quais são responsáveis, dessa forma, são sempre alvos de questionamentos negativos e críticas.

As teorias normativa e construtivista, na maioria dos casos, são incapazes de esclarecer as dimensões produtiva e discursiva originadas pela relação entre recompensas e dívidas intelectuais e os fenômenos persuasivos que embasam as análises, respectivamente. Apesar da incapacidade de se sustentar completamente para todos os questionamentos e abordagens, a teoria normativa apresenta maior consistência em relação à teoria construtivista, quando consideradas em relação às dimensões analíticas de observação e análise dos objetos enquanto fenômenos, originalmente e socialmente construídos no domínio ciência (BORNMANN; DANIEL, 2008).

É necessário destacar também os contextos sociais e culturais relacionados aos estudos de citação, no âmbito da Comunicação Científica. Para isso, Leydesdorff e Wouters (1999) ressaltam a existência de um elo entre texto e contexto, formalizado pela união entre produção, citação e referência que, por um lado, contribuem e impulsionam, porém, por outro, dificultam e limitam a compreensão sobre aspectos que envolvem os diferentes fenômenos abordados pelas teorias vigentes.

Na opinião de Riviera (2013), as práticas de citação utilizadas pelos pesquisadores são influenciadas pelos contextos vivenciados pelos próprios pesquisadores. Isso ocorre, principalmente, quando há a necessidade de convencer, ignorar, refutar ou fundamentar sobre um determinado assunto, ou então quando impõem ou manipulam elementos simbólicos dos quais possuem ou exercem controle ou poder científico. Diante deste contexto, o conjunto de

práticas relacionadas à produção e comunicação do conhecimento torna-se suscetível às circunstâncias do momento, das correntes teórico metodológicas vigentes e das posições ideológicas de grupos e segmentos sociais no campo da ciência (SILVEIRA; CAREGNATO, 2018).

É importante destacar a concepção sociocultural dos estudos no contexto das citações. Diante disso, Silveira e Caregnato (2018, p.58), afirmam que a concepção sociocultural dos estudos de citação

> está orientada para evidenciar e explicar as influências que os múltiplos contextos exercem nas relações existentes entre produção e citação, bem como a repercussão dessa influência para os campos e domínios científicos. Os objetos de análise desta concepção são os contextos culturais direcionados para os processos que constituem as conexões entre as práticas de citação e de produção científica. A identificação e a análise da lógica de reprodução social baseada nas formas de consagração no universo da ciência é o objetivo central desta concepção teórica, entendendo que as relações de força e de poder exercidos pelos atores, grupos e instituições interferem nas formas de citar e produzir (SILVEIRA e CAREGNATO, 2018, p.58).

Reconhecer contribuições de distintos campos e domínios é um contrato que rege termos legais e éticos, fornecendo fácil acesso às outras áreas de pesquisa e facilitando a interoperabilidade com outras temáticas.

Esta seção buscou explanar os conceitos entorno da citação, bem como a apresentação de abordagens históricas e opiniões distintas entre diferentes autores no escopo da citação. Na próxima seção, será abordado sobre o modelo para detecção e extração de referências bibliográficas, de modo a apresentar os detalhes do desenvolvimento, metodologicamente falando, dificuldades e o problema de fato, que o presente trabalho se dispõe a trabalhar.

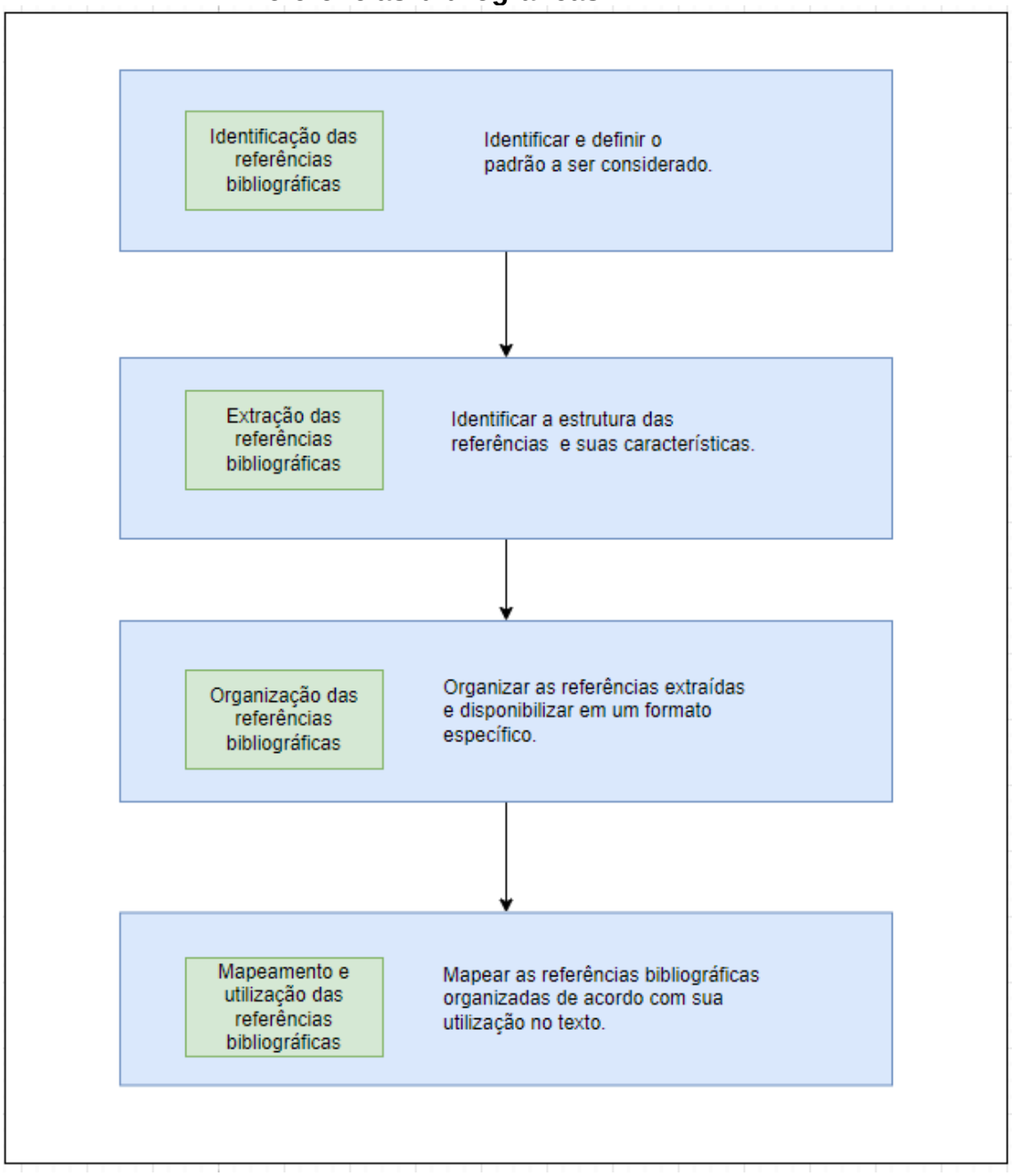
5. MODELO PARA IDENTIFICAÇÃO, EXTRAÇÃO E ORGANIZAÇÃO DE REFERÊNCIAS E CITAÇÕES BIBLIOGRÁFICAS

Esta seção apresenta o detalhamento técnico, estrutura metodológica, problemas, restrições e limitações do modelo proposto. Para comprovar a eficácia do modelo desenvolvido nesta pesquisa, fora implementado um protótipo, apresentado na seção 6, como prova de conceito, que é capaz de identificar, extrair e organizar as citações e referências bibliográficas de trabalhos acadêmicos que estejam dentro das normas ABNT.

Com o desenvolvimento da pesquisa e consequentemente, a busca por parâmetros que justificassem o principal problema deste estudo, observou-se que há inúmeros problemas similares que poderiam ser resolvidos com a mesma metodologia, como as diferentes normas de citação e referência. Entretanto, devido à complexidade de modelar soluções para tais finalidades, esta pesquisa visou abordar apenas um dos problemas, dentre os diversos contidos em seu universo, o qual inclui trabalhos que utilizam referências bibliográficas, com base nas normas ABNT.

O modelo proposto possui a função de estabelecer o fluxo e normas para atingir o resultado final, tal modelo é primordial nesta pesquisa, pois é por meio dele que se pretende atingir aos objetivos descritos e responder a problemática apresentada. Diante do exposto, o modelo construído pode ser visualizado na Figura 4.

Figura 4: Modelo para identificação, extração e organização de referências bibliográficas



Fonte: Elaborado pelo autor (2022).

5.1 Primeira etapa – Identificação das referências bibliográficas

A primeira etapa, que busca identificar as referências bibliográficas, é o ponto de partida. Nesta etapa, houve a necessidade de estipular previamente os padrões considerados na modelagem inicial. Para isso, inicialmente, foram abordados três padrões técnicos:

- ABNT - caracteriza-se pelas normas oficiais da Associação Brasileira de Normas Técnicas, utilizado para referências bibliográficas em textos científicos e acadêmicos;
- APA (*American Psychological Association*) - geralmente utilizado na literatura internacional; e
- MLA (*Modern Language Association)* - comumente mais aplicado em textos das Ciências Humanas e Artes.

Diante dos três padrões técnicos apresentados, a presente pesquisa considera apenas o padrão ABNT, regido pela Associação Brasileira de Normas Técnicas para uso em textos científicos e acadêmicos, de modo a considerar normas voltadas para professores e profissionais que buscam especificações sobre os princípios para a elaboração de trabalhos acadêmicos, como teses, dissertações e trabalhos de conclusão de cursos.

De acordo com a própria Associação Brasileira de Normas Técnicas (1989), seu significado diz respeito ao Foro Nacional de Normalização, de modo que as Normas Brasileiras são de responsabilidade dos Comitês Brasileiros (ABNT/CB), dos Organismos de Normalização Setorial (ABNT/ONS) e das Comissões de Estudos Especiais (ABNT/CEE), em que são desenvolvidas por Comissões de Estudos (CE), formadas por todos interessados em temas voltados ao objeto de normalização. O uso de suas normas é dado voluntariamente, isto é, não há uma obrigatoriedade para se utilizar estas normas. Porém, caso haja alguma expressão explícita em um instrumento do Poder Público, como lei, portaria ou normativa, o uso da ABNT torna-se obrigatório, como é o caso do desenvolvimento de trabalhos científicos em grande parte das instituições no Brasil.

Henriques e Medeiros (2017) relatam ainda que apesar das críticas e discordâncias, as normas ABNT, no Brasil, servem como parâmetro oficial e obrigatório para todos os envolvidos em atividades científicas, técnicas ou acadêmicas. Possuem caráter internacional e estão em vigor nos meios técnicos, científicos e acadêmicos de todos os países do mundo, de modo que não há como ignorá-las, mesmo discordando ou criticando as falhas existentes em sua elaboração.

As referências bibliográficas são compostas por elementos imprescindíveis que se associam à uma função ou papel, tais como: autor,

título da obra, subtítulo (quando houver), edição, local, editora e data de publicação. Há também, a presença de elementos complementares: ilustrador, índice, tradutor, organizador, coordenador, descrição física do volume da obra (número de páginas ou número de volumes, dimensões - em centímetros: largura e altura), série ou coleção, ISBN (*International Standard Book Number*) e notas especiais, como mimeografados, no prelo, não publicado ou título original. Os autores relatam ainda que, nos trabalhos acadêmicos, costumeiramente são indicados apenas os elementos essenciais, tendo sua bibliografia baseada na maioria dos casos em livros, capítulos de livros, dissertação de mestrado, teses de doutorado, revistas, jornais, textos legislativos, arquivos eletrônicos e filmes de vídeo (HENRIQUES E MEDEIROS, 2017).

Para a elaboração do modelo, foram consideradas as normas ABNT com o intuito de fixar a ordem dos elementos bibliográficos e estabelecer convenções para a transcrição e apresentação dos elementos das referências bibliográficas. Neste contexto, há diferentes normas das quais propõem que a referência pode aparecer inteiramente incluída no texto, em parte do texto, parte em nota de rodapé, fim do texto, lista de referências (sinal ético ou analítico) ou no início de resumos ou recensões.

Diante disso, foram consideradas as referências que aparecem no final do texto, na lista de referências. Ressalta-se que, no que tange às citações, suas diferentes formas de aparição não foram explanadas de maneira mais aprofundada, pois elas foram consideradas para uma etapa mais específica do modelo, em que é abordado sobre o mapeamento e uso das referências e citações bibliográficas extraídas em um contexto mais genérico e simplificado.

5.2 Segunda etapa - Extração das referências bibliográficas

Na segunda etapa foi analisado o padrão ABNT, escolhido para iniciar a elaboração do modelo, bem como para o desenvolvimento do protótipo. Para isso, foi analisada a estrutura das referências bibliográfica e seus elementos.

Foram considerados os elementos imprescindíveis das referências bibliográficas, de modo que a estrutura e apresentação de cada elemento são apresentadas a seguir:

- **Autor**: SOBRENOME em letras maiúsculas, vírgula, nome do autor com as iniciais em maiúsculas ou de maneira abreviada, apenas as iniciais, ponto. Caso os nomes sejam abreviados (menos comum), deve-se manter a uniformidade de tratamento para todas as referências:

  ALVES, Primeiro Nome. Título da obra. 1. ed. São Paulo: Atlas, 2005.

  Caso tenham dois ou três autores, deve-se separar por ponto e vírgula; caso haja mais de três autores, após o primeiro referenciado, acrescenta-se a expressão latina et al. (e outros), sem destaque:

  ALVES, Primeiro Nome; ROSA, Segundo Nome; SANTOS, Terceiro Nome. Título da obra. 1. ed. São Paulo: Atlas, 2005.

  Referência com mais de três autores:

  ALVES, Primeiro Nome et al. Título da obra. 1. ed. São Paulo: Atlas, 2005.

  Caso o nome do autor compareça em várias obras, o mesmo deve ser substituído por um traço equivalente a seis espaços, seguido de ponto:

  _____. Título da obra. 1. ed. São Paulo: Atlas, 2005.

  Se tiver um organizador ou coordenador presente na obra, a abreviatura da função deve aparecer entre parênteses:

  ALVES, Primeiro Nome (Org). Título da obra. 1. ed. São Paulo: Atlas, 2005.

  ALVES, Primeiro Nome (Coord). Título da obra. 1. ed. São Paulo: Atlas, 2005.

  De acordo com as últimas normas, este formato foi desconsiderado.

- **Título da obra**: itálico, grifado ou sublinhado (uso incomum), ponto. Na presença de subtítulo, deve ser antecedido de dois-pontos, sem destaque (deve seguir a fonte padrão do texto). Neste caso, considera-se destacado o emprego de qualquer tipo de fonte diferenciada: *bold*, itálica ou outro tipo de escrita. Caso o pesquisador opte pelo tipo itálico,

deve usá-lo em todas as referências; se optou pelo *bold*, igualmente, e assim para todas as variações, conforme exemplo a seguir:

SOBRENOME, Nome. Título da obra: subtítulo da obra. São Paulo: Atlas, 2015.

SOBRENOME, Nome. **Título da obra**: subtítulo da obra. São Paulo: Atlas, 2015.

SOBRENOME, Nome. <u>Título da obra</u>: subtítulo da obra. São Paulo: Atlas, 2015.

- **Subtítulo da obra**: o subtítulo não deve estar destacado, e deve ser antecedido por dois pontos:

  SOBRENOME, Nome. Título da obra: subtítulo da obra. São Paulo: Atlas, 2016.

- **Edição**: informa a edição a partir da segunda, em números arábicos, sem ordinal e com a palavra edição de forma abreviada: 2. ed.:

  SOBRENOME, Nome. Título principal da obra: subtítulo da obra. 2. ed. São Paulo: Atlas, 2012.

- **Local da publicação**: define o nome da cidade onde a obra foi publicada. Neste caso, não pode haver abreviações. Se existirem cidades com o mesmo nome em estados ou países diferentes, anota-se o Estado ou país, seguido por dois-pontos:

  SOBRENOME, Primeiro Nome. Título da obra. Cidade: Nome da Editora, 2005.

  SOBRENOME, Primeiro Nome. Título da obra. Estado: Cidade, Nome da Editora, 2005.

- **Editora**: define o nome da editora após os dois-pontos, sem a razão social, parentescos, etc. Em algumas ocasiões, têm-se as abreviaturas: FGV (Fundação Getúlio Vargas), Edusp (Editora da Universidade de São Paulo), Difel (Difusão Europeia do Livro):

  Cidade: Editora

  Cidade: Nome da Editora

  Caso não haja o local para inserir, deve-se empregar a notação S.1, a letra S em maiúsculo indica o nome de uma cidade, que, evidentemente, seria escrita com a inicial

maiúscula: SOBRENOME, Nome. Título da obra: subtítulo da obra. [S.1.]: Editora, 2000.

Emprega-se também a notação [S.N] (*sine nomine*) para indicar a ausência do editor.

Caso duas editoras sejam as mesmas de um único local:
Cidade: Primeira Editora: Segunda Editora

Caso duas editoras sejam de locais diferentes:
Cidade: Primeira Editora; Cidade: Segunda Editora.

- **Número de volumes**: deve ser indicado com a notação volume de maneira abreviada (v). Caso o número anteceda a notação (2 v.), significa a quantidade de volumes, se o número for inserido depois da notação (v.2), significa o índice numérico do volume, no caso, volume dois.
- **Data**: indica o ano da publicação. Deve-se grafar com algarismos arábicos, sem ponto no milhar, antecedido de vírgula e seguido de ponto. Caso não seja possível identificar a data, anota-se a data aproximada entre colchetes:

SOBRENOME, Primeiro Nome. Título da obra. Cidade: Editora, 1996.

Referência com data provável:

SOBRENOME, Primeiro Nome. Título da obra. Cidade: Editora, [1996?]

Referência com data certa, mas não indicada (recuperada de uma apresentação ou prefácio): [1995]

Referência com data aproximada: [ca. 2001]

Referência com década certa: [199-]

Referência com década provável: [199-?]

Referência com século certo: [19--]

Referência com século provável: [19--?]

Ressalta-se que para todos os elementos apresentados, a margem deve ser feita com alinhamento à esquerda, mantendo um espaço interlinear entre uma referência e outra. Um exemplo de referência com o fluxo técnico detalhado pode ser visto a seguir:

MARTIN, Robert C. **Código limpo**: Habilidades práticas do Agile Software. [S.1]: Alta Books, 2009.

Neste exemplo, tem-se o sobrenome em letras maiúsculas (MARTIN), seguido por vírgula acompanhado de espaço e o nome do autor em caixa alta e baixa, seguido por ponto. Posteriormente, é apresentado o título da obra, em tipo *bold*, com a inicial maiúscula, seguido por ponto. O número da edição deve aparecer somente a partir da segunda, seguido por ponto. Salienta-se que a abreviatura de segunda (2.), e edição (ed), devem seguir os padrões apresentados, respectivamente. Neste caso, não há, pois se encontra na primeira edição. Após o número da edição, há a presença do local, que, neste caso, está rotulado com a notação S.1, o qual indica a ausência do local, seguido por dois pontos. Posteriormente, tem-se o nome da editora, acompanhada por vírgula e, sem a presença do rótulo “Editora”. Por fim, o ano da publicação é apresentado, seguido por ponto final.

Diante das diferentes normas do padrão ABNT, que são atualizadas constantemente, espera-se que a solução embasada neste modelo abarque um escopo mais genérico, de modo que pequenas mudanças não gerem impacto negativo nos resultados. Ademais, espera-se que a solução contemple a norma mais atualizada, mas não descarte as mais antigas.

5.3 Terceira etapa – Organização das referências e citações bibliográficas

Na terceira etapa, os dados processados são organizados de modo que possam ser utilizados na próxima etapa, em que será abordado sobre o mapeamento dos dados obtidos em relação às referências bibliográficas.

Neste ponto, o modelo requer um resultado que inclua todas as referências bibliográficas dos documentos processados em um conjunto de informações, porém, não tratadas. O objetivo é percorrer por todos os grupos de dados extraídos e incluir cada valor em um campo específico. Para tanto, ao final desta etapa, espera-se uma lista de dados, embasados nos elementos imprescindíveis apresentados na seção 5.2, com os seguintes valores explicitados no quadro 2:

Quadro 2: Elementos-chave das referências considerados pelo modelo

Atributo	Valor	Descrição
Autores	Lista de texto	Lista com os autores encontrados no material científico.
Título	Texto	Título do material referenciado.
Informações sobre a edição	Texto	Edição do material referenciado (caso exista).
Ano	Texto	Ano do material referenciado.
Citado por	Inteiro	Quantidade de vezes que os autores foram citados.
Quadrante	Lista	Informa a quantidade de vezes que o autor foi citado em cada quadrante (4 disponíveis).

Fonte: Elaborado pelo autor (2021).

Observa-se que, desta maneira, os valores estão distribuídos e separados, facilitando o consumo e uso para qualquer atividade que possua as referências e citações como objetos de estudo. Independente da tecnologia para implementar essa rotina, espera-se que os resultados estejam disponíveis, posteriormente, em formatos acessíveis e serializáveis, como JSON ou XML.

A escolha do formato é importante, pois a comunicação entre máquinas, mais especificamente entre sistemas web, necessita de um protocolo conhecido por ambos os lados; para que seja feita a integração entre os sistemas. Diante disso, o formato JSON desempenha uma função fundamental nesse quesito, conhecido de maneira universal, é uma linguagem que pode ser facilmente interpretada por desenvolvedores e máquinas, de modo que se tornou o mais popular para efetuar a comunicação entre sistemas que utilizam os protocolos HTTP e HTTPS PEZOA (2016).

Pezoa (2016) discorre ainda que, mediante a popularidade do formato JSON, sua utilização foi considerada em diversos cenários, como em casos de APIs públicas. Entretanto, o autor apresenta alguns problemas em sua utilização, como a falta de uma especificação formal, isto é, o que acontece, por exemplo, se ao invés de enviar uma cadeia de caracteres no JSON, for enviado um número ou até mesmo um valor booleano. Para tanto, caso não haja o apoio de uma camada de integridade, deve-se utilizar uma definição de esquema para pré-estabelecer as solicitações que aderem ao padrão desejado, funcionando como um filtro.

Diante disso, a importância de disponibilizar os dados em formato acessível para outras máquinas está nas prioridades do modelo, para que tais dados possam ser consumidos por outras soluções. A figura 5 demonstra a forma de como devem ser disponibilizados os dados.

Figura 5: Estrutura de dados requerida pelo modelo


```
{
  "autores": [
    "ALLARD, S.",
    "QAYYUM, A.",
    "MEHRA, B. "
  ],
  "titulo": "Intercultural leadership for information
    professionals:  building awareness to effectively serve
    diverse multicultural populations.",
  "info_edicao": "Education Libraries,  v. 30, n. 1,",
  "ano": "2007",
  "disponivel": "None",
  "citadoPor": 1,
  "quadrante": [
    "Terceiro quadrante"
  ]
}
```


Fonte: Elaborado pelo autor (2022).

O atributo ‘Autores’ é especificado como lista de *string* (ou texto), pois, devido à possibilidade de existir mais de um autor responsável pela obra científica, o modelo deverá expor cada registro separadamente dentro da lista, iniciando pelo sobrenome do autor, de modo que facilite a utilização independente de cada registro.

Em seguida, tem-se o ‘Título’ do material científico, tal atributo possui um tipo denominado *string* ou texto, ou seja, o seu conteúdo enquadra-se como uma cadeia de caracteres que armazenam dados textuais.

Posteriormente, apresenta-se a ‘edição’ do material em questão, onde todas as informações relacionadas à edição também estarão disponíveis em formato de *string* (ou texto).

O próximo atributo é o ‘Ano’, neste caso, apesar de que o ano possui números, o seu conteúdo em si representa um texto, logo, optou-se pelo tipo texto para seu mapeamento. Justifica-se também a escolha do tipo texto para o ano, pois, em alguns casos, o ano vem acompanhado de um caractere nas citações, como por exemplo: 2020a). Caso o atributo fosse tratado como numérico, esta situação ocasionaria problemas técnicos.

Depois, apresenta-se o atributo ‘Quantidade de citações’, caracterizado e mapeado como tipo inteiro, seu valor corresponde à quantidade de vezes que o registro foi citado no texto.

Por fim, tem-se o atributo ‘Quadrante’, para cumprir seu propósito, o texto deverá ser dividido em quatro quadrantes. Em seguida, o atributo será mapeado como uma lista de objetos estáticos, inicialmente estabelecidos com quatro valores: (1) Quadrante 1: 0; (2) Quadrante 2: 0; (3) Quadrante 3: 0; e (4) Quadrante 4: 0. Diante disso, tem-se uma ideia da parte em que o autor está sendo citado. Embora não seja possível mapear qual a seção específica com esta abordagem, será possível identificar, com certa precisão, a aparição dos autores em um contexto geral, como no início, meio ou fim do texto. Este atributo corresponde à quantidade de vezes que um determinado autor foi citado dentro de cada quadrante.

5.4 Quarta etapa – Mapeamento e uso das referências e citações tratadas

Nesta etapa, tem-se o resultado tratado e armazenado em uma lista de informações com cada referência extraída do texto, de modo que o consumo das informações seja uma tarefa fácil, ou seja, o conjunto de informações obtido na etapa anterior deve apresentar uma clareza em relação ao seu conteúdo.

Diante disso, necessita-se identificar a quantidade de vezes que um autor foi citado no texto. Para isso, inicialmente faz-se necessário obter o primeiro autor da lista de cada referência e, posteriormente, calcular a quantidade de vezes que uma determinada palavra aparece em um texto. Diante disso, deve-se percorrer cada resultado dentro da lista e executar uma verificação com o nome e/ou sobrenome do autor, de modo que toda vez que encontre algum valor similar no texto, seja incrementado o contador de citações por autor.

Espera-se também que seja possível identificar em qual etapa do texto o autor foi citado. Para isso, com o texto dividido em quatro quadrantes, é feita uma busca com o nome do autor no texto e posteriormente, obtém-se o índice referente à sua posição. Com isso, calcula-se o total de caracteres contidos no texto e identifica-se o quadrante ao qual o resultado da busca pertence.

Para a modelagem deste processo, deve-se considerar algumas limitações e desafios, que encadeiam erros semânticos, abordados na próxima subseção.

5.5 Limitações do modelo

Ressalta-se que o protótipo implementado depende do padrão ABNT apresentado na descrição do modelo nesta pesquisa, isto é, caso o título de um trabalho científico apareça na frente do nome do autor, o modelo projetado apresentará o título como nome do autor, gerando um erro semântico.

Houve a necessidade de definir qual a responsabilidade do modelo em situações em que houver erros de digitação nas referências. Por exemplo, caso uma determinada referência estiver com um caractere espaço em um lugar indevido, o modelo poderia apresentar duas propostas: (i): pausar o processo e apresentar a referência bem como o lugar exato de onde está o erro; ou (ii): desconsiderar o erro de digitação e continuar com o procedimento. Optou-se pela segunda opção, por conta de evitar travamentos no processo, consequentemente tornando-o mais fluído. Ressalta-se que em casos de erros semânticos, como a inversão da ordem de um elemento imprescindível, o modelo não é capaz de detectar inconsistência nesses casos.

Observa-se que determinados padrões de citações não se enquadram mais nas normas atuais, logo, o modelo é projetado para considerar alguns casos, como nas situações em que o nome do autor aparece em várias obras, o mesmo deve ser substituído por um traço equivalente a seis espaços, seguido de ponto, conforme segue:

_____. Título da obra. 1. ed. São Paulo: Atlas, 2005.

Entretanto, apesar de tal abordagem estar fora das normas atuais da ABNT, bem como de suas estruturas mais novas, foi necessário considerá-la para englobar trabalhos que fizeram uso deste padrão.

Com a lista de autores mapeada, é possível desenvolver as próximas funções, referente ao índice de citações de cada autor e suas respectivas aparições no texto. Para isso, deve-se considerar problemas como:

- Nomes ou sobrenomes de autores iguais;
- Nomes ou sobrenomes digitados erroneamente.

Por fim, a efetividade dos resultados depende do detalhamento do modelo que realiza a identificação e extração das referências, mas, não apenas isso, espera-se também que as referências sejam digitadas de maneira correta, considerando o padrão ABNT de citação e referência.

6. IMPLEMENTAÇÃO DO PROTÓTIPO

Para a implementação do protótipo, foi utilizada a linguagem de programação Python, devido à sua facilidade por trabalhar com dados em diferentes volumes.

A linguagem Python foi concebida a partir de outra linguagem denominada ABC, Python foi criado por Guido van Rossum, em 1940, no Instituto Nacional de Pesquisa para Matemática e Ciência da Computação da Holanda (CWI), e focava originalmente em usuários como físicos e engenheiros. Python é uma linguagem interpretada, possui licença aberta de código (*open source*) e é compatível com a *General Public License* (GPL), porém, com menos restrições, permitindo sua inclusão em produtos proprietários. Ademais, a linguagem possui uma sintaxe clara e concisa, favorecendo a legibilidade do código-fonte, tornando a linguagem mais produtiva (BORGES, 2014).

Para o desenvolvimento do motor de processamento, é necessário explicar sobre a principal metodologia para exercer tal função; as expressões regulares. Optou-se pelo uso de expressões regulares no protótipo diante da necessidade de se trabalhar com detecção de padrões e processamento de dados. Nesse contexto, Friedl (2006) relata que o bom uso de expressões regulares libera poderes de processamento que ajudam a solucionar grandes e pequenos problemas quais são pequenos, mas poderiam ser transformados em grandes se não fosse o uso de expressões regulares.

Entretanto, é sabido que existem alguns problemas para se trabalhar com elas, como o tempo de execução em volumes altos, podendo impactar negativamente na performance do protótipo.

Historicamente, Friedl (2006) disserta que as sementes das expressões regulares foram plantadas no início da década de 1940 por dois neurofisiologistas: Warren McCulloch e Walter Pitts, os quais desenvolveram, de acordo com suas teorias, modelos de trabalho sobre o sistema nervoso no nível do neurônio. Ainda de acordo com Friedl (2006), as expressões regulares tornaram-se realidade muitos anos depois, quando Stephen Kleene, matemático, descreveu formalmente esses modelos em álgebra, denominados conjuntos regulares, que, posteriormente, foram concebidos pelo autor por meio de uma notação simples, nomeando-os para expressões regulares.

As expressões regulares foram estudadas desde sua origem e ao longo das décadas de 1950 e 1960, inclusive na teorética dos círculos matemáticos. Apesar de existirem evidências de estudos anteriores que abordaram sobre expressões regulares, o primeiro uso computacional publicado sobre o tema foi desenvolvido por Ken Thompson, em 1968, chamado '*Regular Expression Search Algorithm'*, responsável pela produção de um compilador de expressão regular Friedl (2006).

Jargas (2012) define uma expressão regular como um método formal de se especificar um padrão de texto. O autor ressalta ainda que, de modo detalhado, pode-se definir também como uma composição de símbolos, caracteres com funções especiais, que, se agrupam entre si com caracteres literais, constituem uma sequência, uma expressão que é interpretada como uma regra que resulta em sucesso; caso uma entrada de dados condiz com o padrão da regra especificada, ou seja, segue exatamente todas as condições impostas.

Algumas variações sobre a definição de uma expressão regular podem ser aceitas, Jargas (2012) apresenta algumas:

- uma maneira de procurar uma cadeia de caracteres que não se sabe exatamente como é, mas há uma ideia de suas possíveis variações;
- um modo de identificar um trecho em posições específicas;
- um modo para especificar padrões complexos que podem ser procurados e condizentes à uma cadeia de caracteres.

Ressalta-se que este protótipo não possui a intenção de ser a versão final do produto tecnológico propriamente dito, mas sim um algoritmo que servirá como prova de conceito para os objetivos estipulados.

O protótipo possui quatro funções que são executadas, respectivamente:

- Identificar as referências bibliográficas;
- Extraí-las;
- Organizá-las;
- Mapear conforme seu uso no texto.

Do ponto de vista do usuário, o fluxo de interação com a proposta desenvolvida é simples, basta fazer o upload dos arquivos para uma pasta específica no computador e executar o programa. A partir desta etapa, todo o fluxo é feito e disponibilizado pela solução implementada.

Diante de todo o aspecto conceitual sobre as variáveis imprescindíveis para referenciar trabalhos científicos no âmbito da ciência, fora implementado um algoritmo capaz de ler documentos (trabalhos científicos) e transformar o conteúdo lido em padrões detectados aderentes aos padrões ABNT, apresentados anteriormente.

6.1 Identificação

O funcionamento do protótipo se dá inicialmente por importar os arquivos em formato .pdf ou .docx para uma pasta específica, cuja leitura e escrita pelo algoritmo é liberada. Posteriormente, uma biblioteca capaz de realizar a leitura de arquivos .pdf e.docx, com base no caminho informado no algoritmo. É possível executar o algoritmo com um ou mais arquivos dentro da pasta, porém, o resultado será entregue em apenas um arquivo. Optou-se pela abordagem de entrega em um único arquivo, em formato JSON, para facilitar a leitura de todas as referências, sem a necessidade de relacionar mais de um arquivo. Entretanto, é possível adaptar o protótipo para separar cada resultado em um arquivo distinto.

Destaca-se que os arquivos para teste devem utilizar referências bibliográficas ABNT e estarem disponíveis (abertos) à leitura, isto é, existem

alguns arquivos que são bloqueados ou em formato de imagem, logo, não serão considerados para os testes.

Com a configuração prévia realizada e, certamente, com o ambiente de desenvolvimento preparado, com a instalação do Python e *runtime* do Java, são declaradas algumas variáveis de controle, apresentadas a seguir.

1. *NumberOfPages;*
2. *FullResponse;*
3. *FullResponseLength;*
4. *ReferenceIndex;*
5. *ResponseToRegex.*

A variável *NumberOfPages* (1) é responsável por armazenar o número de páginas. Inicialmente, foi utilizada para calcular o índice da página em que apareciam as referências. A princípio, este processo funcionava da seguinte forma, calculava-se o total de páginas, criava-se um laço com base no índice 0 e o total de páginas do arquivo com extensão .pdf, e, em cada iteração, era armazenado o conteúdo lido em uma variável. Com isso, duas variáveis foram alteradas, a primeira, responsável pelo conteúdo total armazenado, e a segunda, o conteúdo atual lido na iteração.

Posteriormente, procurava-se pela palavra ‘referências’ somente na variável que armazenava o conteúdo da iteração, e, caso fosse encontrado, era criada uma variável denominada *ReferencePageIndex*, responsável por guardar o último índice da página em que a palavra referências fosse encontrada. Ressalta-se que a variável com o conteúdo atual da iteração e a variável que armazena o índice de página da referência - são sobrescritas a cada repetição. Com o índice inicial da página que estavam as referências, um novo laço era feito para obter o conteúdo em texto das referências, logo, utilizava-se o último índice armazenado na variável *ReferencePageIndex* como base inicial e o total de páginas, como índice final. Em cada iteração do laço, era armazenado o conteúdo na variável ‘*ResponseToRegex’* para ser consumido nas etapas posteriores.

Esta abordagem obteve êxito em todos os casos, porém, caso a palavra referências fosse encontrada entre as referências ou após, em outras seções,

as etapas posteriores da execução eram impactadas de maneira negativa. Entretanto, é sabido que tomando a palavra ‘referências’ como ponto de partida não engloba todos os casos, uma vez que, em situações onde as referências estejam em uma seção com outra terminologia, ocasionará em uma falha.

Outro ponto a se considerar foi a performance do algoritmo, com o aumento do número de páginas ou de trabalhos inseridos na fila de execução, o tempo para leitura dos arquivos era mais alto devido às rotinas utilizadas para obter o conteúdo.

Logo, esta abordagem foi desconsiderada, com a troca deste processo, foi alterada a biblioteca de leitura de arquivos com extensão .pdf, utilizou-se uma outra abordagem diferente da dinâmica de laços e páginas, explicada no detalhamento da variável *FullResponse*. Esta etapa é primordial para o bom funcionamento do protótipo, de modo a gerar resultados com alta taxa de sucesso, pois, segundo Chenet (2017), caso a obtenção do texto ocorra muito cedo, a informação pode vir inconsistente. Por outro prisma, se o conteúdo estiver muito volumoso e com informações fora de contexto, no caso, que não condizem com as referências bibliográficas, haverá muito ruído e aumento de complexidade nas próximas etapas.

A variável *FullResponse* (2) possui a função de armazenar o conteúdo total do arquivo. Para isso, a biblioteca utilizada para ler o conteúdo de arquivos entrega toda a resposta de uma única vez, o que certamente deixa o código mais performático, uma vez que, não é necessário construir um laço para obter o conteúdo de cada página, conforme descrito na abordagem anterior. A biblioteca denomina-se como *Tika* e possui funções como detecção de tipo de documento e extração de conteúdo em diferentes formatos. Para efetuar sua instalação, é necessário ter o *runtime* da linguagem de programação Java instalado na máquina, pois o código da biblioteca é desenvolvido em Java.

A próxima variável é a *FullResponseLength* (3), que possui a função de armazenar o tamanho do conteúdo total (variável anterior, 2). Para isso, utiliza-se uma função em Python chamada ‘*len’*, que recebe uma *string* como parâmetro, isto posto, o conteúdo total, seu retorno é um número referente à quantidade de caracteres do parâmetro recebido.

A posteriori, foi declarada a variável *ReferenceIndex* (4), responsável por obter o índice da palavra ‘referências’ no conteúdo total. Para isso, utilizou-se

uma função chamada ‘*rindex*’, que obtém o índice da última ocorrência de uma determinada palavra. Nesta situação, pode ocorrer um problema, caso a palavra ‘referências’ seja encontrada entre ou após a seção de referências. Porém, durante os testes, não foram encontradas situações que tivessem casos como esse.

Por fim, a variável *ResponseToRegex* (5), armazena o conteúdo consumido pela expressão regular, por meio de uma função que obtém um intervalo de caracteres do conteúdo total, iniciando no índice disponível na variável *ReferenceIndex* com o adicional de 11 (quantidade de caracteres na palavra ‘referências’) e finalizando no índice presente na variável *FullResponseLength*. O adicional de onze no último índice de referência se dá mediante a necessidade de remover a própria palavra do conteúdo a ser aplicado à expressão regular.

A figura 6 apresenta a declaração inicial de variáveis que compõem a implementação do protótipo.

Figura 6: Declaração inicial das variáveis do protótipo

```
for files in read_files:
    raw = parser.from_file(files)
    fullResponse = raw['content']
    fullResponseLength = len(fullResponse)
    referenceIndex = fullResponse.lower().rindex("referências")
    responseToRegex = fullResponse[referenceIndex+11:fullResponseLength]
    fullResponse = fullResponse[0:referenceIndex]
```

Fonte: Elaborado pelo autor (2022).

É importante ressaltar que todo o conteúdo armazenado para uso posterior, foi convertido para minúsculo por meio da função *lower*, com o intuito de padronizar os dados para consumo da expressão regular, evitando desencontros e falhas de identificação.

Pode-se observar também que, a variável *FullResponse* (2) passa por um outro processo no final, ou seja, ela é sobrescrita pelo seu próprio intervalo de caracteres entre o índice 0 (inicial) e o índice da variável 4. Isso ocorre para deixar separado o texto das referências, com o fito de melhorar a qualidade dos resultados e segregar as responsabilidades para que a etapa de mapeamento possa ser realizada com eficiência.

Observa-se, ainda que, as variáveis de armazenamento responsáveis por gerenciar o conteúdo total e o conteúdo a ser consumido pela expressão regular estão em formato texto. Talvez haja a possibilidade de aplicar a expressão regular diretamente no arquivo .pdf ou .docx, porém, foi considerado efetuar a conversão do arquivo em texto, para melhores resultados. Em nível de leitura dos arquivos, existem diversas bibliotecas disponíveis gratuitamente na linguagem Python. Inicialmente, foi testada a biblioteca PyPDF2, porém, com a mudança de abordagem explicada na variável 1, a alteração do mecanismo fora necessária.

6.2 Extração

Conforme o detalhamento do modelo, foram consideradas as referências no padrão ABNT, logo, na etapa atual foi necessário definir os meios técnicos para realizar a extração das referências. Para isso, optou-se por trabalhar com expressões regulares, devido a sua flexibilidade e dinamicidade para obter dados em diferentes formatos. Ademais, pode-se dizer que grande parte da manutenção e evolução da solução se dá em alterações na própria expressão regular.

Com o conteúdo do material científico presente na variável de armazenamento *ResponseToRegex*, a qual obtém o conteúdo referente ao texto, exceto a lista de referências, inicia-se o processo de submeter o conteúdo da variável a um processamento via expressão regular que detecta os padrões estabelecidos nesta pesquisa, de modo que todo resultado encontrado vá para uma cadeia de resultados compatíveis separados em grupos com funções distintas, sendo autores, título da obra, editora, ano e local de disponibilidade do material.

Extração dos autores contidos nas referências bibliográficas

De acordo com os dados fornecidos pelo modelo, os primeiros dados considerados para extração foram os autores, para isso, foi criado um grupo dentro da expressão regular que possui os seguintes requisitos:

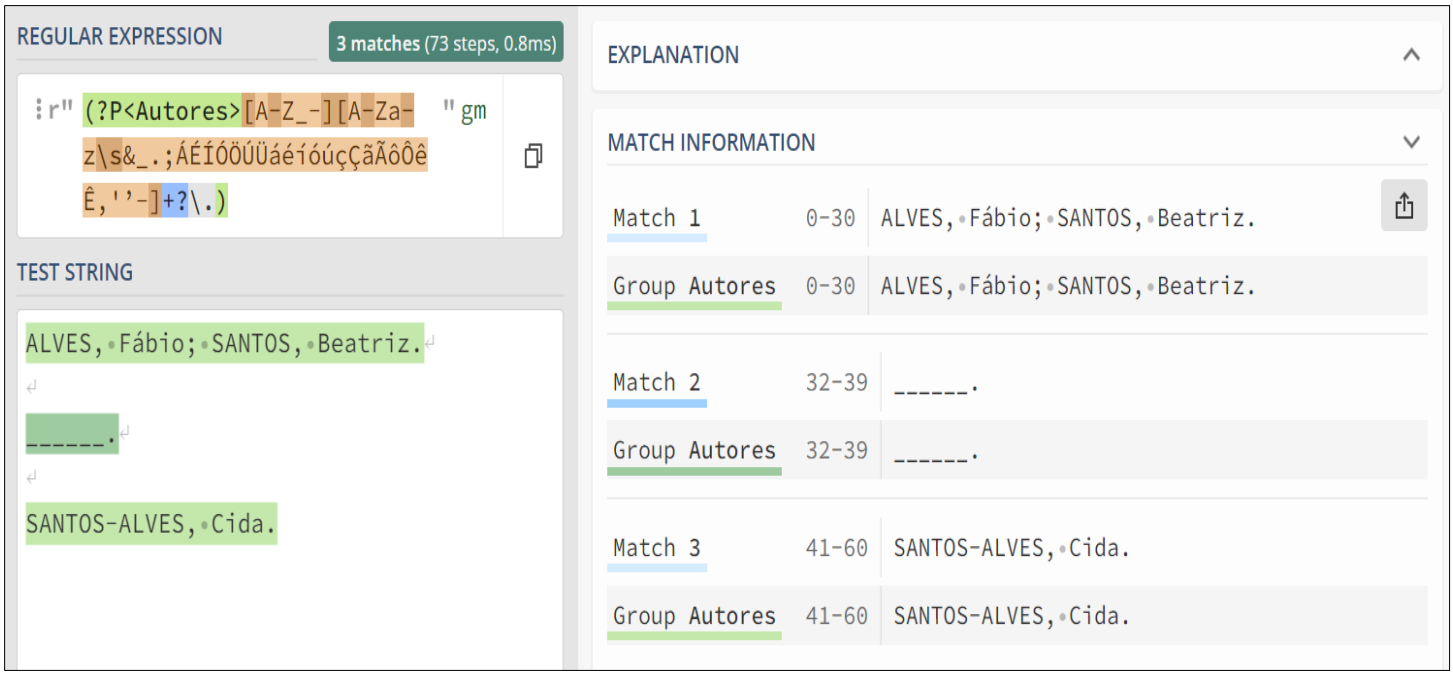
- Uma sequência de caracteres em letras maiúsculas que inclui letras de A-Z, traço, acentuações e um caractere denominado *underline*.
- Um traço equivalente a seis espaços (opcional);
- Uma vírgula (,);

- Uma sequência de caracteres iniciado por uma letra maiúscula;
- Um espaço opcional;
- Um ponto final.

Em versões anteriores dos padrões ABNT é possível encontrar um traço que equivale a seis espaços caso o nome do autor apareça em diferentes obras, logo, foi necessário acrescentá-lo à regra para incluir trabalhos que aderissem às normas anteriores da ABNT.

A expressão regular apresentada na figura 7 demonstra a forma técnica utilizada para obter os autores.

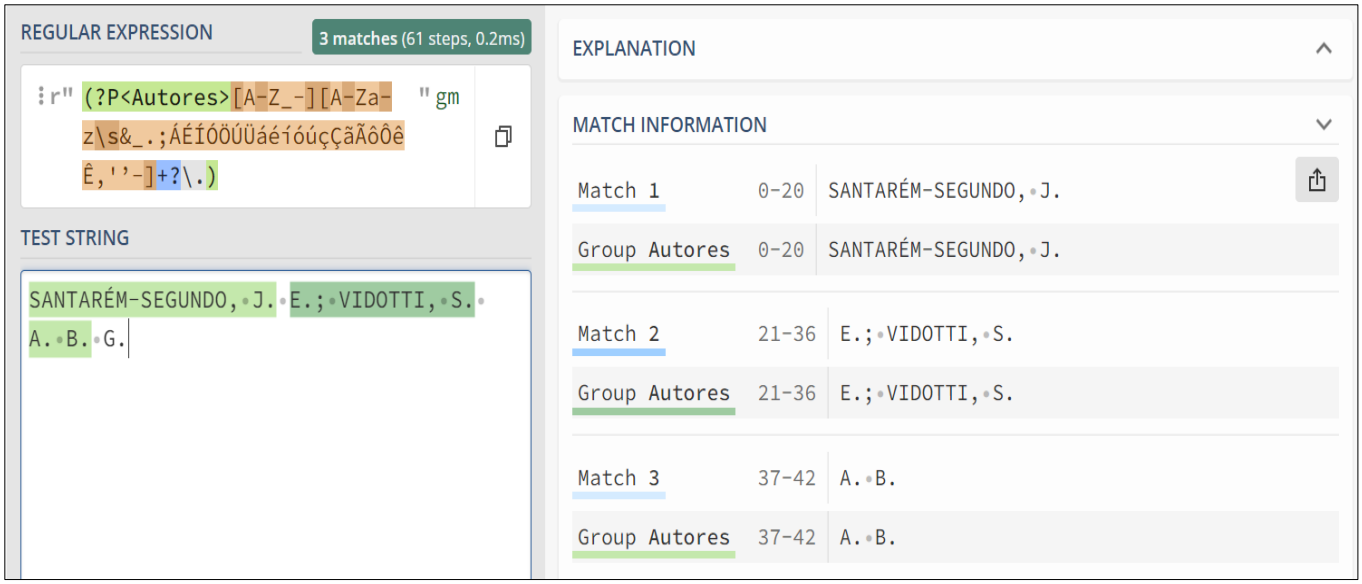
Figura 7: Simulação de extração de autores com expressão regular



Fonte: Elaborado pelo autor (2022).

Para efetuar os casos de testes, foi utilizada a plataforma Regex101, que permite simular ocorrências de texto baseado em expressões regulares. Em alguns casos, o sobrenome do autor possui mais de uma abreviação, o que acarreta problemas na expressão regular acima, conforme apresenta a figura 7.

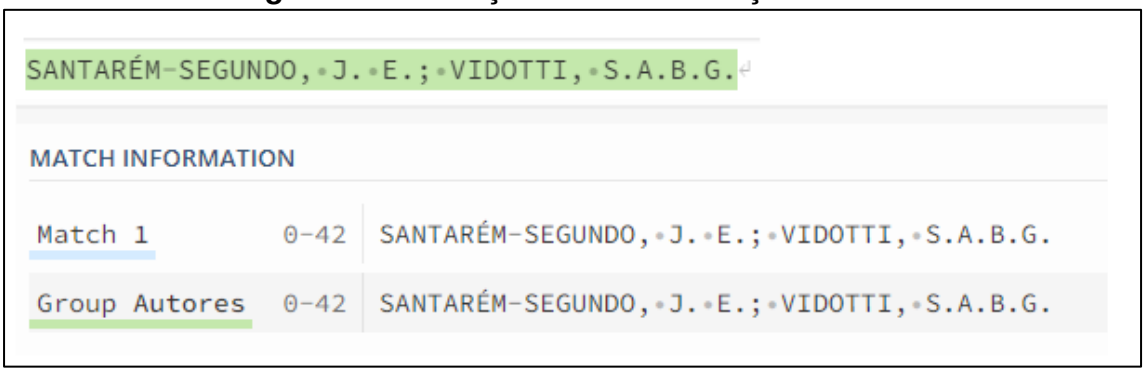
Figura 8: Simulação de extração de autores após adaptação da expressão regular



Fonte: Elaborado pelo autor (2022).

Observa-se que em ambos os casos, o resultado está incorreto, pois as abreviaturas do nome não foram extraídas, gerando resultados inconsistentes, logo, foi necessário fazer um ajuste na expressão regular para corrigir esses casos, o que impactou em problemas nos casos anteriores. Diante disso, foi necessário adaptar a expressão regular para considerar esses casos e corrigir os anteriores a nível de código, isto é, com tratativas independentes de expressão regular, pode-se observar o resultado na figura 9.

Figura 9: Simulação final da extração de autores



Fonte: Elaborado pelo autor (2022).

Com a adaptação desenvolvida na expressão regular foi possível obter o nome dos autores com mais eficiência, porém, para os casos que possuem apenas uma abreviatura, parte do título é incorporado ao nome do autor, ocasionando inconsistência. Para resolver isso, foi implementada uma rotina que obtém o excesso adicional incorreto no nome do autor e transfere para o título. Nota-se também que os autores são separados por ponto e vírgula (;), isso ocorre porque a expressão regular não foi implementada para entregar os resultados em forma de lista, logo, esta alteração foi feita utilizando outras abordagens, detalhadas na seção de organização.

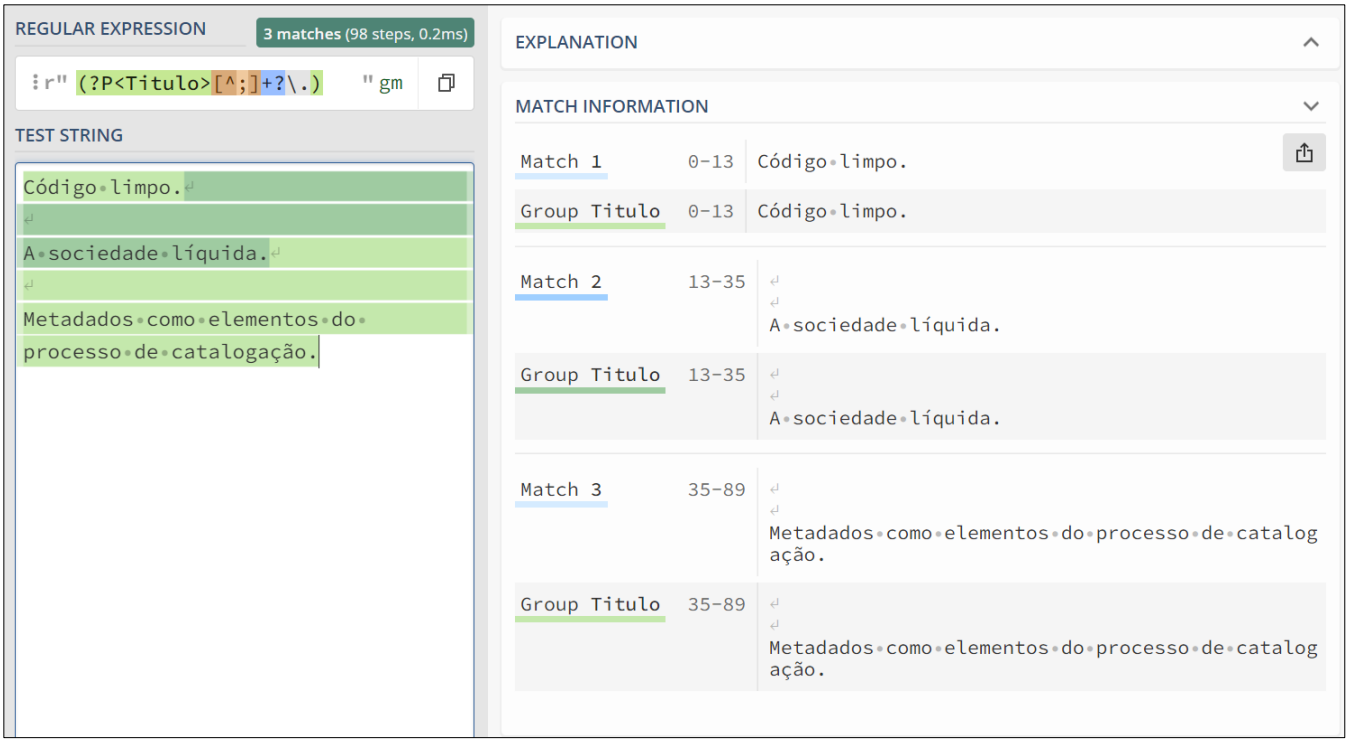
Extração do título

O próximo registro a ser extraído é o título do material científico, para isso, criou-se um grupo com as seguintes normas:

- Uma sequência de caracteres iniciado por letra maiúscula, que pode incluir acentuações, dois pontos e ponto e vírgula;
- Ponto final.

A expressão regular desenvolvida, bem como os resultados do teste, pode ser visualizada na figura 10.

Figura 10: Simulação de extração de título



Fonte: Elaborado pelo autor (2022).

Ressalta-se que o título pode finalizar com qualquer pontuação, além do ponto final, que é o mais comum. A expressão regular desenvolvida permite que o título inicie com qualquer caractere, isto é, número, acentuação, letra minúscula, etc. Optou-se por abranger qualquer ocasião para obter o máximo de sucesso possível. É importante destacar novamente que o protótipo não possui a intenção de informar erros nas referências, mas sim contorná-los com o objetivo de extrair a informação mais coerente possível.

Extração de informações relacionadas à edição

Para extrair as informações voltadas para a edição foram necessárias algumas regras que incluem letras, números, abreviações e acentuações. O funcionamento efetivo desta parte da expressão regular se dá somente se as partes anteriores forem acopladas a ela, devido a forma como foi construída a expressão regular de modo geral.

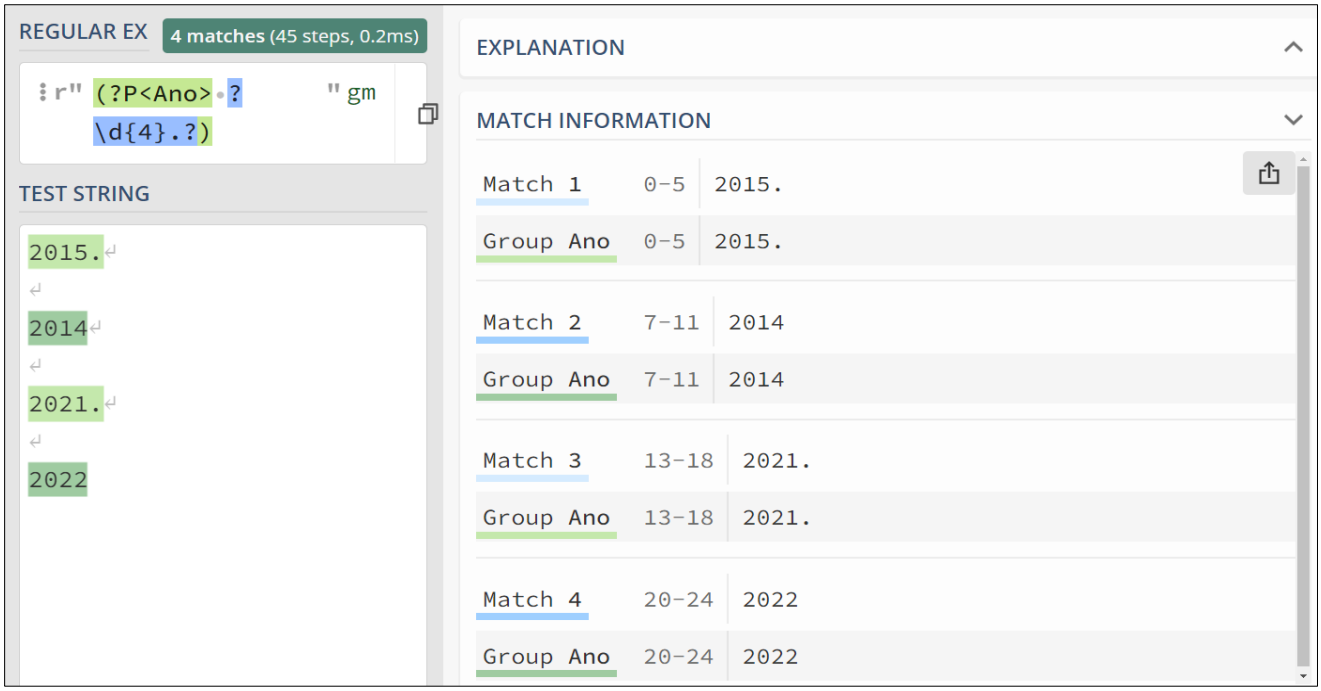
Ressalta-se que as informações da edição estão previstas para serem obtidas no terceiro bloco, ou seja, caso haja alguma outra informação após o título, será considerado informação da edição até que se encontre um ponto final seguido por quatro dígitos, indicando o ano da publicação do material.

Extração do ano

Posteriormente, tem-se a extração do ano da publicação, para isso, foram considerados 4 caracteres numéricos finalizados por um ponto final opcional, conforme segue a expressão regular apresentada na figura 11.

(?P<Ano> ?\d{4}.?)

Figura 11: Simulação de extração de ano



Fonte: Elaborado pelo autor (2022).

Optou-se por colocar o ponto final como opcional para abranger casos em que esteja faltando devido a alguma falha de escrita. Outro ponto a se considerar é que o conteúdo que antecede ou sucede o ano é indiferente, a expressão regular é capaz de detectar em meio a qualquer informação, o que garante uma maior efetividade quando inserida junto com as expressões anteriores.

6.3 Organização

Os valores considerados para esta foram embasados nos elementos imprescindíveis apontados na seção 5.2. A inclusão de novos valores depende da modificação do modelo, sendo necessário reconhecer o padrão no texto, passando por todas as condições e tratativas necessárias para o processo.

A etapa de organização é responsável por adaptar todo o conteúdo processado pela expressão regular, realizando as devidas formatações, removendo espaços e adaptando formatos.

Para isso, o primeiro passo foi migrar o formato dos autores para uma lista de texto, logo, foi utilizada uma função para separar conjuntos de palavras com base em um caractere específico, no caso, o ponto e vírgula, conforme apresenta o exemplo a seguir:

ALVES, Beatriz Pinheiro; BARATA, Germana.

Com a aplicação da função ‘split’, tem-se o seguinte resultado:

['ALVES, Beatriz Pinheiro', ' BARATA, Germana.']

Isso ocorre pois facilita a utilização individual de cada autor na etapa de mapeamento, além de gerar uma divisão mais coerente dos dados extraídos.

Em relação ao título, informações da edição e ano, esses dados são submetidos a uma função de reparação, onde são removidos espaços adicionais e caracteres codificados. Ao final da execução dessas rotinas, tem-se o resultado com uma lista de objetos que armazena cada referência. O resultado foi transformado em formato JSON, conforme apresenta a figura 12.

Figura 12: Resultado contendo as referências bibliográficas tratadas


```
[
  {
    "autores": [
      "ARAÚJO, Vânia M.R.H. "
    ],
    "titulo": "Sistemas de recuperação da informação: nova abordagem teórico conceitual.",
    "informacoes_edicao": "Ciência da Informação, Brasília, v. 24, n. 1",
    "ano": "1995",
    "disponivel": " Disponível em: < > ",
    "quantidadeCitacoes": 1,
    "quadrante": [
      "Primeiro quadrante"
    ]
  },
  {
    "autores": [
      "BAEZA-YATES, R.",
      "RIBEIRO-NETO, B. "
    ],
    "titulo": "Modern information retrieval.",
    "informacoes_edicao": "New York : ACM,",
    "ano": "1999",
    "disponivel": "",
    "quantidadeCitacoes": 1,
    "quadrante": [
      "Primeiro quadrante"
    ]
  },
  {
    "autores": [
      "CRONIN, Blaise. "
    ],
    "titulo": "Esquemas conceituais e estratégicos para a gerência da informação.",
    "informacoes_edicao": "Revista da Escola da Biblioteconomia da UFMG, Belo Horizonte, v. 19, n. 2, p. 195-220.",
    "ano": "1990",
    "disponivel": "",
    "quantidadeCitacoes": 1,
    "quadrante": [
      "Quarto quadrante"
    ]
  }
]
```


Fonte: Elaborado pelo autor (2022).

A adoção do formato JSON (*JavaScript Object Notation*), se deve pelos seguintes motivos:

- Formato aberto, flexível e amplamente utilizado;
- Permite a transferência e consumo de dados por protocolos HTTP entre diversas aplicações computacionais de maneira rápida e consistente;
- Apresenta uma estrutura organizada e de fácil compreensão;
- Aceito por diversas ferramentas que trabalham com análises de dados.

Em determinados casos podem existir informações valores indevidos, como espaços ao final de alguns valores, pontuações entre outros formatos. As inconsistências verificadas, como espaços ao final dos valores, pontuações indevidas, expressões com espaços, entre outras, podem ser tratadas com

base em ajustes no modelo e na expressão regular do protótipo, que é o motor de processamento da solução proposta ou então com ajustes a nível de código, como a função reparadora que processou os dados contidos no título, informações da edição e o ano.

A I4OC (2022) promove uma iniciativa para a disponibilização de dados sobre citações estruturadas, separáveis e abertas. Pode-se observar que a organização e o resultado desta etapa do protótipo condizem com os objetivos dessa iniciativa, disponibilizando os dados de referências estruturáveis e separáveis.

6.4 Mapeamento

Na etapa atual, é feito o mapeamento dos dados correspondentes ao modelo, ou seja, identificar a quantidade de vezes que os autores são citados no texto e, posteriormente, extrair a parte do texto em que está sendo citado.

Diante disso, a identificação da quantidade de vezes em que um autor foi citado no texto foi implementada inicialmente, por meio de uma busca simples, onde é comparado o sobrenome do primeiro autor da lista com o conteúdo do texto, de modo que para cada resultado encontrado, incrementa-se uma variável controladora.

Porém, isso desencadeou dois problemas:

- Nomes duplicados que não condizem com o autor;
- Nomes duplicados de autores com ano de publicação divergente.

Para contornar esses problemas, foi implementada uma nova expressão regular que filtra pelo sobrenome do autor juntamente com o ano de publicação, logo, foi possível extrair os dados que realmente correspondem aos nomes dos autores, além da capacidade de diferenciar autores iguais com publicações em anos diferentes.

Por fim, a identificação da parte em que o autor é citado requer a divisão do texto científico em quatro partes, para isso, foi necessário obter o tamanho do texto e dividi-lo por quatro. Ressalta-se que o tamanho do texto, neste caso, não considera a seção de referências. Por exemplo, considerando que o material científico possua um tamanho equivalente a 62225 caracteres, pode-se considerar os seguintes quadrantes, sendo o primeiro valor, o índice inicial e o segundo, o índice final.

- Primeiro quadrante: 0 – 15.556
- Segundo quadrante: 15.556 – 31.112
- Terceiro quadrante: 31112 – 46.668
- Quarto quadrante: 46.668 – 62.224

6.5 Validação

Para validar o modelo proposto, foi realizado um teste com base em um conjunto de 10 artigos científicos retirados do Google Scholar e Periódicos Capes utilizando-se os seguintes termos: comunicação científica; web semântica; citação e tecnologia. Foram considerados trabalhos que estavam no padrão ABNT de referências e citações. A tabela 1 demonstra os resultados obtidos.

Tabela 1: Resultados do protótipo

Trabalho	Quantidade de Referências	Identificadas	Extraídas com sucesso	Extraídas parcialmente	Taxa de sucesso
Artigo 01	11	11	10	1	90,91%
Artigo 02	11	11	11	0	100%
Artigo 03	21	21	20	1	95,24%
Artigo 04	52	49	49	0	94,23%
Artigo 05	30	30	29	2	96,67%
Artigo 06	12	10	10	0	91,67%
Artigo 07	4	4	4	0	100%
Artigo 08	18	18	17	1	94,44%
Artigo 09	118	116	109	7	92,37%
Artigo 10	23	21	21	0	91,30%
Total	**325**	**269**	**174**	**95**	**94,68%**

Fonte: Elaborada pelo autor (2021).

Pode-se observar que em alguns artigos a taxa de sucesso é maior, isso ocorre, pois, as referências identificadas são mais aderentes ao modelo proposto nesta pesquisa. Em outros casos, onde a taxa de sucesso é menor, ocorreram problemas na identificação das referências devido às limitações do modelo apresentadas na subseção 5.5. Ressalta-se que o protótipo identificou textos que não eram propriamente referências, porém, devido ao seu formato de expressão muito similar ao modelo proposto, acabou influenciando negativamente nos resultados. A taxa de sucesso foi calculada considerando a quantidade total de referências em cada trabalho. Desta forma, considerando

que a taxa fosse calculada somente sobre as referências identificadas, haveria um aumento positivo na taxa de sucesso, uma vez que, as referências que não foram identificadas no início deixariam de entrar no cálculo da taxa de sucesso.

As extrações parciais não foram contabilizadas na taxa de sucesso, pois, em alguns casos foi possível extrair somente algumas informações, como o nome dos autores e o título do trabalho apenas, o restante das informações foi extraído indevidamente devido à forma com que estavam escritas ou por inconsistências no protótipo implementado.

Outro ponto importante para se considerar é que alguns trabalhos apresentaram referências que possuem estruturas similares, porém em grupos diferentes, o que desencadeia uma extração inconsistente. Isso ocorre devido a forma de trabalho da expressão regular, que, de maneira simplória, realiza o reconhecimento de padrões, ou seja, caso haja algum trecho do texto que se constitui de modo similar à uma referência, este texto será considerado no processamento.

Diante disso, tais trabalhos foram descartados para a análise. Houve também problemas na leitura de arquivos com codificações distintas, isto é, arquivos que não são passíveis de leitura por meio da tecnologia utilizada, incapacitando o protótipo implementado de obter as informações contidas no arquivo.

Apesar da solução técnica permitir a execução de vários arquivos de uma única vez, os testes foram feitos separadamente, contabilizando os resultados de modo individual.

Espera-se que com a resolução dos desafios e limitações apresentadas, o protótipo consiga entregar os resultados mais consistentes e adaptar-se à complexidade no contexto das referências e citações.

6.6 Limitações

O protótipo utiliza expressão regular como base principal para o motor de extração. Apesar de existirem algumas tratativas e adaptações que transcendem a responsabilidade da expressão regular, considera-se que melhores resultados podem ser obtidos somente com alterações nas normas de reconhecimento de padrões da expressão regular utilizada.

A atual expressão regular possui algumas limitações, ou seja, ela não é capaz de lidar com as seguintes situações:

- Identificar referências que estão sem o ano de publicação;
- Diferenciar abordagens similares, mas em grupos diferentes, isto é, existem trechos do texto muito similares à uma expressão regular, mas não é de fato uma expressão regular;
- Obter todos os nomes e sobrenomes de autores sem a necessidade de adaptações a nível de código.

A obtenção dos nomes e sobrenomes dos autores se dá por meio de uma expressão regular que inclui parte do título no nome do autor, ocasionando uma extração incorreta, com excesso de caracteres adicionais no nome dos autores. Isso ocorreu devido a uma adaptação feita para abranger nomes que incluíssem múltiplas abreviações. Diante disso, o excesso adicional incorreto é retirado por uma rotina de adaptação desenvolvida somente para sanar esses casos.

A nível geral, o protótipo não é capaz de reconhecer arquivos .pdf em formato de imagem, o que influencia diretamente na leitura do conteúdo. Entretanto, é sabido que existem soluções para reconhecer esses casos, logo, considera-se como uma evolução significativa.

7 CONSIDERAÇÕES FINAIS

O presente estudo partiu do seguinte problema: como é possível criar indicadores quantitativos ou realizar análises de citações e referências bibliográficas em segmentos gerais ou específicos? Com base nesse problema, objetivou-se a constituição de um modelo do processamento de informações que tornasse possível a identificação, extração e a organização de referências e citações de textos em língua portuguesa, com base nas normas ABNT. Nesse contexto, pode-se afirmar que o problema de pesquisa foi respondido no decorrer do estudo, tanto a partir da análise de referencial teórico quanto com base no desenvolvimento do modelo que foi proposto no objetivo, que, portanto, também fora alcançado. Além disso, é possível inferir que, ao considerar a natureza interdisciplinar da Ciência da Informação, abre-se margem para outras áreas do conhecimento, com a finalidade de contribuir em pesquisas como esta, isto é, áreas que possuem arcabouço teórico voltado à ciência, comunicação científica e teoria das citações.

Como pode ser notado no referencial teórico deste estudo, existem inúmeros estudos de citação no campo da ciência que visam identificar a qualidade de uma determinada obra - com base no número de referências que ela utiliza ou até mesmo pela quantidade de vezes que tal obra foi citada em outros trabalhos. Dessa maneira, esta pesquisa buscou levantar questões importantes acerca deste tema e teve como foco implementar, um modelo capaz de identificar e extrair referências e citações de textos em língua portuguesa no padrão ABNT.

Ressalta-se que existem diferentes tecnologias para implementar a solução presente neste estudo, e isso tende a aumentar, considerando a evolução da tecnologia. Com o aumento de publicações científicas nos mais diversos campos do conhecimento, soluções como esta se tornam pertinentes dado a necessidade de análises de referências e citações, visto que tal tarefa é complexa, vagarosa e passível de erros quando realizada manualmente.

Algumas iniciativas permitem fazer este trabalho de maneira automática, mas isso ocorre para padrões específicos ou o modelo é acoplado à solução técnica, que muitas vezes não estão acessíveis às normas brasileiras de padronização, mais especificamente as da ABNT. Diante disso, cria-se um

cenário específico para aplicar esta solução, que pode ser evoluída para considerar outros padrões de normalização acadêmica.

Ademais, como melhorias futuras, considera-se que trabalhar com técnicas de *machine learning* ou até mesmo outros tipos de algoritmos embasados em expressão regular - podem ser consideradas opções promissoras para evolução da solução desenvolvida nesta dissertação. Logo, acredita-se que seja viável a implementação com outras tecnologias para validar a taxa de sucesso. Outrossim, este trabalho amplia as possibilidades de estudos sobre evolução de pesquisa, mapeamento de produção em periódicos e eventos científicos no Brasil.

Além disso, o modelo e, consequentemente, o protótipo, não são capazes de identificar outros tipos de padrões, ou seja, é esperado o padrão ABNT para que haja sucesso na execução. Tem-se como evolução a identificação automática de padrões, de modo que seja indiferente para os resultados.

Ressalta-se ainda que o modelo e o protótipo construídos possam sofrer alterações para melhorar os resultados. Ainda assim, é possível afirmar que o objetivo geral do estudo foi contemplado, assim como a problemática explorada e, a princípio, solucionada, uma vez que a pesquisa conseguiu avançar de maneira significativa no que tange ao processo de identificar, extrair e organizar citações e referências bibliográficas, ampliando e facilitando os estudos de citação, uma vez que, automatiza-se o processo e gera resultados tratados para serem utilizados em diferentes abordagens, especialmente pelos pesquisadores que estudam a produção e organização do conhecimento.

REFERÊNCIAS

AKEN, Joan E. van. Management research based on the paradigm of the design sciences: the quest for field-tested and grounded technological rules. **Journal of management studies**, v. 41, n. 2, p. 219-246, 2004. Disponível em https://www.researchgate.net/publication/4868922_Management_research_based_on_the_paradigm_of_the_design_sciences_The_quest_for_field-tested_and_grounded_technological_Rules. Acesso em 20 fev. 2021.

ALPAYDIN, Ethem. Introduction to machine learning. MIT press, 2014.

ALVARENGA, Lídia. Bibliometria e arqueologia do saber de Michel Foucault: traços de identidade teórico-metodológica. **Ciência da informação**, v. 27, p. 253-261, 1998. Disponível em http://revista.ibict.br/ciinf/article/view/778. Acesso em 10 fev. 2021.

ALVES, R. C. V**. Metadados como elementos do processo de catalogação**. ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. Normas para citações e referências Bibliográficas. São Paulo: ABNT, 2003.

BARATA, Germana. Em revisão: o impacto da produção científica brasileira para o Brasil. **Ciência e Cultura**, v. 67, n. 4, p. 06-08, 2015.

BARBOSA, D. M., BAX, M. A Design Science como metodologia para a criação de um modelo de Gestão da Informação para o contexto da avaliação de cursos de graduação. **Revista Ibero-Americana De Ciência Da Informação**, v.10, n.1*,* 32–48, 2017. Disponível em https://periodicos.unb.br/index.php/RICI/article/view/2471. Acesso em 20 mar. 2021.

BAUMAN, Zygmunt. A sociedade líquida. **Folha de São Paulo**, v. 19, p. 4-9, 2003.

BERNAL, John Desmond et al. The social function of science. **The Social Function of Science.** 1939.

BORGES, Luiz Eduardo. **Python para desenvolvedores: aborda Python 3.3**. Novatec Editora, 2014.

BORNMANN, L.; DANIEL, H.-D. What do citation counts measure? A review of studies on citing behavior. **Journal of Documentation**, v. 64, n. 1, p. 45- 79, 2008. Disponível em https://www.emerald.com/insight/content/doi/10.1108/00220410810844150/full/html. Acesso em 29 mar. 2021.

BRAGA, Gilda Maria. Informação, Ciência, Política Científica: o pensamento de Derek de Solla Price. **Ciência da Informação**, Brasília, DF, v. 3, n. 2, p. 155-177, 1974. Disponível em http://revista.ibict.br/ciinf/article/view/50. Acesso em 17 fev. 2021.

BAX, M. P. Design Science: filosofia da pesquisa em Ciência da Informação e tecnologia. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15. 2014. Belo Horizonte. **Anais...** Belo Horizonte: UFMG, Programa de Pós-Graduação em Ciência da Informação, 2014.

BUNGE, M. **Ciência e desenvolvimento.** Belo Horizonte: Itatiaia; São Paulo: EDUSP, 1980. (O Homem e a Ciência, v. 11).

CARIBÉ, Rita de Cássia do Vale. Comunicação científica: reflexões sobre o conceito. **Informação & Sociedade: Estudos; v. 25, n. 3 (2015); 89-104**, v. 24, n. 2, p. 104-89, 2015. Disponível em https://periodicos.ufpb.br/ojs/index.php/ies/article/view/23109. Acesso em 16 set. 2021.

CARVALHO, Maria Martha de. Análises Bibliométricas da Literatura de Química no Brasil. **Ciência da Informação**, Brasília, DF, v. 4, n. 2, p. 119-141, 1975. Disponível em http://revista.ibict.br/ciinf/article/view/56. Acesso em 19 set. 2021.

COZZENS, S.E. What do citations count? The rhetoric-first model. **Scientometrics***.* 1989, vol. 15, nº 5–6, pp. 437–447. Disponível em https://link.springer.com/article/10.1007/BF02017064. Acesso em 11 fev. 2021.

DAS GRAÇAS TARGINO, Maria. Comunicação científica: uma revisão de seus elementos básicos. **Informação & Sociedade**, v. 10, n. 2, 2000. Disponível em https://periodicos.ufpb.br/ojs/index.php/ies/article/view/326. Acesso em 10 set. 2021.

DEMO, P. **Metodologia do conhecimento científico**. São Paulo: Atlas, 2000.

DRESCH, A.; LACERDA, D. P.; ANTUNES JUNIOR, J. A. V. **Design Science Research:** A Method for Science and Technology Advancement. New York: [s.n.], 2015.

ERIKSON, M. G.; ERLANDSON, P. A taxonomy of motives to city. **Social Studies of Science**, v. 44, n. 1, p. 1-13, 2014. Disponível em https://journals.sagepub.com/doi/abs/10.1177/0306312714522871. Acesso em 10 ago. 2021.

FOUREZ, G; ROUANET, L.P; FOUREZ, G. **A construção das ciências**. Unesp, 1995.

FRAGA DO AMARAL E SILVA, E. **Um sistema de extração de informação em referências bibliográficas baseado em aprendizagem e máquina**. 2004. Dissertação de Mestrado. Universidade Federal de Pernambuco.

FRIEDL, Jeffrey E. F. **Mastering Regular Expressions**. 3ª. Edição. O'Reilly Media Inc. Sebastopol. EUA. 2006.

GARVEY, W. D. **Communication:** the essence of science. New York: Pergarnon Press, 1970, p. IX. [apud Hernánez- Canadas, 1987]

GIL, A.C. **Como elaborar projetos de pesquisa**. 4.ed. São Paulo: Atlas, 2009. 175p.

GILBERT, G.N. The transformation of research findings into scientific knowledge. **Soc. Stud. Sci.** 1976, vol. 6, nº 3/ 4, pp. 281–306. Disponível em https://journals.sagepub.com/doi/abs/10.1177/030631277600600302?journalCode=sssb. Acesso em 19 out. 2021.

GRÁCIO, M.C.C. Acoplamento bibliográfico e análise de cocitação: revisão teórico-conceitual. **Encontros Bibli:** revista eletrônica de biblioteconomia e ciência da informação, v.21, n.47, p.82-99, set./dez., 2016. Disponível em https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2016v21n47p82. Acesso em 11 jun.2022.

HENRIQUES, Antonio; MEDEIROS, João Bosco. **Metodologia Científica da Pesquisa Jurídica.** 09/2017. São Paulo. Atlas. Disponível em: https://www.grupogen.com.br/e-book-metodologia-cientifica-da-pesquisa-juridica.

HERNÁNDEZ-CANADAS, Patricia Liset. **Os periódicos "Ciência Hoje" e "Ciência e Cultura" e a Divulgação da Ciência no Brasil**. Rio de Janeiro: ECA! UFRJ! IBICT, 1987.

HJORLAND, B. Citation analysis: A social and dynamic approach to knowledge organization. **Information Processing and Management**, v.49, n.6, p.1313-1325, 2013. Disponível em https://www.sciencedirect.com/science/article/abs/pii/S0306457313000733. Acesso em 11 mar. 2022.

HOFFNAGEL, Judith C. A prática de citação em trabalhos acadêmicos. **Cadernos de Linguagem e Sociedade**, v. 10, n. 1, p. 71-88, 2009. Disponível em https://periodicos.unb.br/index.php/les/article/view/9277. Acesso em 22 set. 2021.

I4OC. **Initiative for Open Citations**, 2022. Disponível em https://i4oc.org/#. Acesso em 11 abr. 2022.

JARGAS, AURELIO MARINHO. **Expressões Regulares: uma abordagem divertida**. Novatec Editora, 2012.

KERLINGER, F. N. **Metodologia da pesquisa em ciências sociais:** um tratamento conceitual. São Paulo: EPU/ EDUSP, 1979.

LALANDE, A**. Vocabulário técnico e crítico da filosofia**. 3. ed. São Paulo: Martins Fontes, 1999.

LETA, Jacqueline. Indicadores de desempenho, ciência brasileira e a cobertura das bases informacionais. **Revista USP**, n. 89, p. 62-67, 2011.

LEYDESDORFF, L. Theories of citation? **Scientometrics**, Amsterdam, v. 43, n. 1, p. 5-25, 1998. Disponível em https://www.researchgate.net/publication/225180439_Theories_of_Citation. Acesso em 10 nov. 2020.

LE COADIC, Y.-F. **A Ciência da Informação**. Brasília: Briquet de Lemos/Livros, 1996. 119 p.

LEYDESDORFF, Loet; WOUTERS, Paul. Between texts and contexts: Advances in theories of citation? (A rejoinder). **Scientometrics**, v. 44, n. 2, p. 169-182, 1999. Disponível em https://link.springer.com/article/10.1007/BF02457378. Acesso em 10 out. 2020.

LEYDESDORFF, L.; AMSTERDAMSKA, O. Dimensions of Citation Analysis. Science, **Technology, & Human Values**, New York, v. 15, n. 3, p. 305-335, 1990. DOI: https://doi.org/10.1177%2F016224399001500303. Disponível em https://journals.sagepub.com/doi/10.1177/016224399001500303. Acesso em 15 ago. 2020.

MARCH, S. T.; SMITH, G. F. Design and natural science research on information technology. **Decision Support Systems**, v. 15, p. 251-266, 1995. Disponível em https://www.sciencedirect.com/science/article/abs/pii/0167923694000412. Acesso em 22 set. 2021.

MARTINELLI, D. P.; VENTURA, C. A. A. (org). **Visão Sistêmica e Administração:** conceitos, metodologias e aplicações. Editora Saraiva, 2005.

MERTON, R.K. **The normative structure of science**. In: MERTON, R.K., ed. The Sociology of Science: Theoretical and Empirical Investigations. Chicago, IL: University of Chicago Press, 1973. pp. 267–278.

MEADOWS, Arthur Jack. **A comunicação científica**. Brasília, DF: Briquet de Lemos, 1999.

MIRANDA, Dely Bezerra. O periódico científico como veículo de comunicação: uma revisão de literatura, **Ci. Inf.** Vol 25(3), 1996, artigos on-line. Disponível em http://revista.ibict.br/ciinf/article/view/4428. Acesso em 10 set. 2021.

MORAES, M.; FURTADO, R. L.; TOMAÉL, M. I. Redes de citação: estudo de rede de pesquisadores a partir da competência em informação. **Em Questão**, Porto Alegre, v. 21, n. 2, p. 181-202, mai/ago 2015. Disponível em https://seer.ufrgs.br/EmQuestao/article/view/47481. Acesso em 19 mar. 2020.

MOREL, Regina Lúcia de Moraes; MOREL, Carlos Médicis. Um Estudo Sobre a Produção Científica Brasileira, Segundo os Dados do Institute for Scientific Information (ISI). **Ciência da Informação, Brasília**, DF, v. 6, n. 2, p. 99-109, 1977. Disponível em http://revista.ibict.br/ciinf/article/view/85. Acesso em 23 ago. 2020.

MUELLER, Suzana Pinheiro Machado. A ciência, o sistema de comunicação científica e a literatura científica. **Fontes de informação para pesquisadores e profissionais. Belo Horizonte: UFMG**, 2000.

NICOLAISEN, J. Citation analysis. **Annual Review of Information Science and Technology**, Baltimore County, v. 41, p. 609-641, 2007. Disponível em https://asistdl.onlinelibrary.wiley.com/doi/10.1002/aris.2007.1440410120. Acesso em 08 mai. 2020.

Normas ABNT. **ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS.**, 2022. Disponível em https://www.normasabnt.org/abnt-nbr-14724/. Acesso em 12 abr. 2022.

PEZOA, Felipe et al. Foundations of JSON schema. In: **Proceedings of the 25th International Conference on World Wide Web**. 2016. p. 263-273.

PRICE, DJ de S. The structures of publication in science and technology. **Factors in the Transfer of Technology**, p. 91-104, 1969.

PRODANOV, FREITAS, **Metodologia do trabalho científico**: Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico. 2.ed.Novo Hamburgo: Feevale, 2013.

ROSENHEAD, J. What's the problem? An introduction to problem structuring methods. **Interfaces**, v. 26, n. 6, p. 117-131, 1996. Disponível em https://www.researchgate.net/publication/247823670_What's_the_Problem_An_Introduction_to_Problem_Structuring_Methods. Acesso em 10 mar. 2021.

ROSSEAU, Ronald. Indicadores Bibliométricos e Econométricos para a Avaliação de Instituições Científicas. **Ciência da Informação**, Brasília, DF, v. 27, n. 2, p. 149- 158, maio/ago. 1998. Disponível em https://www.scielo.br/j/ci/a/7XV6zyFf7bpVYtchQqbtnCd/abstract/?lang=pt. Acesso em 10 nov. 2020.

SANTA ANNA, J. **COMUNICAÇÃO CIENTÍFICA E O PAPEL DOS PERIÓDICOS CIENTÍFICOS NO DESENVOLVIMENTO DAS CIÊNCIAS**. 2019. Disponível em: https://periodicos.ufpb.br/index.php/biblio/article/view/44365. Acesso em 19 fev. 2021.

SANTARÉM-SEGUNDO, J. E.; VIDOTTI, S. A. B. G. Organização da informação na web: a busca na qualidade do armazenamento e da recuperação com a utilização de XML e RDF. In: SIMPÓSIO EM FILOSOFIA E CIÊNCIAS, 5. 2003, Marília. **Anais...** Marília: Unesp Marília Publicações, 2003.

SANTOS, Antonio Raimundo. **Metodologia científica:** a construção do conhecimento. 7ª. Lamparina, 01 de agosto de 2015.

SANTOS, P. L. V. A. C.; ALVES, R. C. V. Metadados e Web Semântica para estruturação da Web 2.0 e Web 3.0. **DataGramaZero**, Rio de Janeiro, v. 10, n. 6, dez. 2009. Disponível em http://www.brapci.inf.br/index.php/article/download/52958. Acesso em 21 ago. 2020.

SCHWARTZMAN, Simon. **Um espaço para a ciência: a formação da comunidade científica no Brasil**. Simon Schwartzman, 2001.

SHANNON, C. E.; WEAVER, W. **The mathematical theory of communication**. Urbana, Ill.: University of Illinois Press, 1949.

SILVEIRA, M. A. A.; CAREGNATO, S. E.; BUFREM, L. S. Estudo das razões das citações na Ciência da Informação: proposta de classificação. **Tendências da Pesquisa Brasileira em Ciência da Informação**, João Pessoa, v. 7, n. 2, p. 232-250, 2014. Disponível em http://hdl.handle.net/20.500.11959/brapci/119535. Acesso em 10 mar. 2019.

SILVEIRA, Murilo Artur Araújo da; CAREGNATO, Sônia Elisa. Demarcações epistemológicas dos estudos de citação: concepção sociocultural das citações. **Perspectivas em Ciência da Informação**, v. 23, p. 55-70, 2018.

DA SILVEIRA, Murilo Artur Araújo; CAREGNATO, Sônia Elisa. Demarcações epistemológicas dos estudos de citação: teorias das citações. **Em Questão**, v. 23, n. 3, p. 250-275, 2017.

SILVA, José Aparecido da; BIANCHI, Maria de Lourdes Pires. Cientometria: a métrica da ciência. **Paidéia (Ribeirão Preto)**, v. 11, p. 5-10, 2001.

SILVEIRA, Tatiana Scalco et al. **Divulgação e política científica:** do bar do mane a ciência hoje (1982-1998). 2000.

SIMON, H. A. The Sciences of the Artificial. 3. ed. Cambridge: MIT Press, 1996.

SOLLA PRICE, Derek J. de. **O Desenvolvimento da Ciência:** análise histórica, filosófica, sociológica e econômica. Rio de Janeiro: Livros Técnicos e Científicos, 1976.

SOLLA PRICE, D. J. **Little science, big science.** New York: Columbia University Press, 1993.

THEOLOGI-GOUTI, P; VITORATOS, E. The role of museum of science in the technological age. **Museologia**, v. 1, p. 67-84, 2001. Disponível em

https://www.researchgate.net/publication/322143810_The_role_of_the_Science_and_Technology_Museum_in_the_development_of_Patras_University_cultural_landscape. Acesso em 23 fev. 2021.

VAISHNAVI, V. K.; KUECHLER, W. **Design science research methods and patterns**. 2. ed. Boca Raton: CRC Press, 2015.

VALENTE, M.E; CAZELLI, S; ALVES, F. Museus, ciência e educação: novos desafios. **História, ciências, saúde-Manguinhos**, v. 12, p. 183-203, 2005. Disponível em https://www.scielo.br/j/hcsm/a/8kBtsgnNggwkjCVYwwFCsGS/?format=pdf&lang=pt. Acesso em 12 abr. 2021.

VAN AKEN, J.E. Management research as a design science: Articulating the research products of mode 2 knowledge production in management. **British journal of management**, v. 16, n. 1, p. 19-36, 2005. Disponível em https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8551.2005.00437.x. Acesso em 19 fev. 2021.

VAN AKEN, J. E. The research design for design science research in management. **Eindhoven:[sn]**, 2011.

VIEIRA, J. G. S. Metodologia de pesquisa científica na prática. **Curitiba: Editora Fael**, 2010.

ZHANG, Q; CAO, Y-G; YU, H. Parsing citations in biomedical articles using conditional random fields. **Computers in biology and medicine**, v. 41, n. 4, p. 190-194, 2011. Disponível em https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3086470/. Acesso em 2 nov. 2020.

ZIMAN, John. **A força do conhecimento**. Belo Horizonte: Itatiaia, 1981. 380p.

ZOU, Jie; LE, Daniel; THOMA, George R. Locating and parsing bibliographic references in HTML medical articles. **International Journal on Document Analysis and Recognition (IJDAR)**, v. 13, n. 2, p. 107-119, 2010. Disponível em https://www.researchgate.net/publication/45272581_Locating_and_parsing_bibliographic_references_in_HTML_medical_articles. Acesso em 19 fev. 2021.