EIXO TEMÁTICO:
Organização e Representação da Informação e do Conhecimento

PRÁTICA TERMINOLÓGICA PARA ATUALIZAÇÃO DE VOCABULÁRIOS CONTROLADOS A PARTIR DA EXTRAÇÃO AUTOMÁTICA DE TERMOS COM O USO DE SOFTWARES

TERMINOLOGICAL PRATICE FOR UPDATING CONTROLLED VOCABULARIES FROM THE AUTOMATIC EXTRACTION OF TERMS USING SOFTWARES

Larissa Kawana Passos Benatto Perandre Pereira[1]
Francisco Carlos Paletta[2]
Marcos Antonio de Moraes[3]

Resumo: A prática terminológica é datada de pelo menos três séculos, advinda das necessidades de denominar conceitos à termos e informações específicas em diversas áreas do conhecimento. Apesar dos procedimentos manuais serem grande parte da metodologia de um trabalho terminológico, esses processos demandam tempo excessivo, especialmente nas etapas iniciais da trajetória do estudo, e diante disso, busca-se por mecanismos tecnológicos mais rápidos e eficientes para otimizar essas etapas. A questão prioritária deste estudo é demonstrar que há necessidade do emprego de ferramentas tecnológicas como parte da metodologia que possa auxiliar nos procedimentos do trabalho terminológico manual sendo fundamental para otimizar essas práticas e obter resultados mais eficientes. O objetivo deste estudo é demonstrar através de uma análise comparativa como as metodologias automatizadas podem ser benéficas nas etapas do processo de trabalho terminológico na atualidade, apontando as características do trabalho automatizado ou semiautomatizado a partir de uma análise teórica de softwares de indexação e de prospecção de dados que fazem mineração textual em documentos de informação. A metodologia utilizada neste estudo será a análise de conteúdo com caráter descritivo. Os resultados apresentam alguns softwares de indexação e mineração de textos, suas características, funcionalidades e aplicações no campo da indexação e representação de documentos informacionais. O estudo conclui que é inevitável o auxílio da tecnologia no trabalho terminológico proporcionando redução de tempo e melhora na produtividade do profissional. Algumas percepções apenas são possíveis por meio das competências humanas, tais como a determinação e atribuição de unidades lexicais

[1] Mestranda em Ciência da Informação pelo Programa de Pós-graduação em Ciência da Informação da Universidade Estadual de Londrina (PPGCI/UEL). E-mail: larissa.benatto@uel.br
[2] Doutor em Ciência da Informação pela Universidade Carlos III de Madrid. Docente da Universidade de São Paulo (USP). Docente do PPGCI/UEL. E-mail: fcpaletta@usp.br
[3] Doutor em Ciência da Informação pelo Programa de Pós-graduação em Ciência da Informação da Universidade Estadual Paulista “Júlio, de Mesquita Filho” (PPGCI/Unesp). Docente do Departamento de Ciência da Informação UEL. E-mail: marcosmoraes@uel.br.

especializadas que possam complementar a representação do documento informacional.

Palavras-chave: Terminologia. Indexação. Automação. Prática terminológica. Vocabulário controlado.

Abstract: Terminological practice dates back at least three centuries, arising from the need to name concepts, terms, and specific information in different areas of knowledge. Although manual procedures are a large part of the methodology of terminological work, these processes require excessive time, especially in the initial stages of the study trajectory, and in view of this, there is a search for faster and more efficient technological mechanisms to optimize these stages. The priority issue of this study is to demonstrate that there is a need to use technological tools as part of the methodology that can assist in manual terminological work procedures, being essential to optimize these practices and obtain more efficient results. The objective of this study is to demonstrate through a comparative analysis how automated methodologies can be beneficial in the stages of the terminological work process today, pointing out the characteristics of automated or semi-automated work based on a theoretical analysis of indexing and prospecting software. data that performs textual mining in information documents. The methodology used in this study will be content analysis with a descriptive nature. The results present some text indexing and mining software, their characteristics, functionalities, and applications in the field of indexing and representation of informational documents. The study concludes that the assistance of technology in terminological work is inevitable, reducing time and improving professional productivity. Some perceptions are only possible through human skills, such as the determination and assignment of specialized lexical units that can complement the representation of the informational document.

Keywords: Terminology. Indexing. Automation. Terminological Practices. Controlled Vocabularies.

1. INTRODUÇÃO

A prática terminológica é datada de pelo menos três séculos, advinda das necessidades de denominar conceitos à termos e informações específicas em diversas áreas do conhecimento, tais como, a química e a biologia. Tornou-se uma prática necessária quando houve avanços da ciência, fazendo com que os especialistas buscassem maneiras de representar a informação que eles obtinham através de suas pesquisas. A partir da primeira metade do século XX, meados dos anos trinta, buscou-se não somente denominar conceitos e suas relações, sobretudo, buscou-se novas denominações de conceitos por meio de um avanço progressista e exponencial das ciências e das tecnologias (Almeida, 2003).

Existem algumas teorias e práticas que são exploradas e abordadas pelos processos terminológicos a exemplo da Teoria Geral da Terminologia (TGT) que tem por características a importância da padronização da terminologia que evita a ambiguidade dos termos, melhorando a comunicação especializada e o uso do termo.

Já a Terminologia Teórica e Aplicada (TTA) tem duas características principais que concentram-se em uma abordagem interdisciplinar para o estudo das unidades lexicais especializadas com foco na criação de vocabulários controlados de baixo grau de complexidade, e a Teoria Comunicativa da Terminologia (TCT), que considera a adaptabilidade e o dinamismo da terminologia conforme as necessidades do usuário, seguindo o preceito de que as unidades lexicais precisam estar de acordo com fatores externos, assim como os sociais e culturais, promovendo uma visão mais abrangente da terminologia.

Apesar dos procedimentos manuais serem grande parte da metodologia de um trabalho terminológico, esses processos demandam tempo excessivo, especialmente nas etapas iniciais da trajetória do estudo. Diante disso, busca-se por mecanismos tecnológicos mais rápidos e eficientes para otimizar essas etapas. A informática e a terminologia são dois campos que se conectam em uma cooperação. Almeida, Oliveira e Aluísio (2006, p. 42) enfatizam que essa junção vem por meio histórico de países desenvolvidos com tradição em terminologia. A ligação às ferramentas automatizadas proporciona a elaboração de bases de dados terminológicos, que aprimoram o processamento da linguagem natural (PLN).

Almeida, Oliveira e Aluísio (2006, p. 42), ressaltam que “A terminologia na era da informática significa criar um conjunto de procedimentos automatizados ou semiautomatizados que deem suporte às tarefas envolvidas no trabalho terminológico”.

As etapas do trabalho terminológico independentemente da maneira como é feito (manual ou automatizado) parte da premissa que esse procedimento é constituinte de etapas, que tem por finalidade realizar o fichamento dos textos a procura de possíveis termos candidatos que farão parte de etiquetas de representação de objetos informacionais de uma base de dados ou até mesmo da construção de um instrumento de controle terminológico - os Sistemas de Organização do Conhecimento (SOC’s) - mas isso depende da necessidade e do objetivo principal de realizar esse estudo especializado. A extração de unidades representativas que são propensas a tornarem-se candidatas quando estão relacionadas ao domínio, e tendo sua definição parte desse contexto específico.

A questão prioritária deste estudo é demonstrar que há necessidade do emprego de ferramentas tecnológicas como parte da metodologia que possa auxiliar nos procedimentos do trabalho terminológico manual sendo fundamental para otimizar

essas práticas e obter resultados mais eficientes, pois através dessas tecnologias é possível reduzir o tempo despendido nas tarefas que envolvem o trabalho terminológico; minimização de erros humanos, e potencializando a padronização dos resultados. Além disso, as ferramentas tecnologias proporcionam a manipulação de um grande volume de dados tendo um acesso mais rápido às informações que se busca.

Portanto, o objetivo deste estudo é demonstrar através de uma análise comparativa como as metodologias automatizadas podem ser benéficas nas etapas do processo de trabalho terminológico na atualidade, apontando as características do trabalho automatizado ou semiautomatizado a partir de uma análise teórica de softwares de indexação e de prospecção de dados que fazem mineração textual em documentos de informação. Para isso, foi realizada uma fundamentação teórica consistente e clara sobre terminologia em Ciência da Informação e depois a apresentação de dois softwares, mediante a análise da literatura publicada sobre cada um deles.

2. AUTOMATIZAÇÃO DO TRABALHO TERMINOLÓGICO NA CIÊNCIA DA INFORMAÇÃO

A terminologia tem caráter teórico-aplicado, que em seu ciclo teórico procura entender como as unidades lexicais são criadas e como se comportam em sua funcionalidade. Para o trabalho aplicado, há delegação de agrupar essas unidades lexicais compondo um Sistema de Organização do Conhecimento (vocabulário controlado especializado), de um ou mais idiomas para uma comunicação de longo alcance. Independentemente de como será realizado o procedimento terminológico, é necessário a identificação da terminologia (temática), e posteriormente, a determinação das unidades lexicais que irão compor o material de referência especializado, atendendo às funcionalidades da obra e necessidades do usuário. (Krieger, 2005, p. 2).

O trabalho terminológico precisa atender os patamares cognitivo, linguístico e discursivo. No âmbito cognitivo é pertinente analisar a temática das unidades lexicais constituintes do vocabulário controlado, isto é, identificar a onomasiologia do termo; referente a averiguar os diferentes conceitos que um termo pode ter. Isso acontece porque um termo pode ser multidisciplinar e conflitante à várias áreas do

conhecimento. Existem dois fatores da composição linguística da unidade lexical especializada, sendo ela a morfossintática que está relacionada à estrutura e função do termo, e a sintagmática que estuda as relações entre os termos, e com essas características juntas, tem-se um vocabulário ampliado. Portanto, o termo é parte poliédrico das teorias metodológicas e para a comunicação, isto é, para que a unidade lexical especializada componha um trabalho terminológico, é necessário que seja comunicativo de forma intrínseca, que para Krieger (2005), “Todos esses focos de análise, quando levados em consideração, representam uma complementaridade analítica de grande apoio ao pesquisador que realiza sua tarefa com fins terminográficos”.

O trabalho terminológico automatizado é um processo fundamental no âmbito da Ciência da Informação, pois através dos recursos informacionais é possível realizar o tratamento da informação de forma mais prática e precisa, facilitando a representação da informação e posteriormente, a recuperação da informação. É uma prática emergente principalmente dentro dessa área do conhecimento, pois através desses mecanismos conseguimos fazer aquilo que era um grande fenômeno depois dos anos 1940: a produção exagerada de informação. As práticas terminológicas semiautomatizadas empregam ferramentas a fim de sistematizar informações sendo suporte das teorias terminológicas.

Com isso, a Teoria Geral da Terminologia (TGT) analisada por Eugen Wüster em 1931 em sua tese de doutorado onde aborda pontos chave da terminologia que vão além da definição do conceito. Dentre esses pontos, inclui-se a precisão e consistência na comunicação científica, significação e estabilidade do termo, coleta e análise para a criação de instrumentos de controle terminológico e a aplicação da terminologia em diferentes áreas do conhecimento. (Cervantes, 2009, p. 126).

> Entretanto, a TGT vem sendo criticada desde os anos de 1990, pois essa teoria não possui recursos suficientes para uma descrição precisa e abrangente do vocabulário técnico. Apesar disso, a TGT é sistemática e coerente, útil para solucionar um dos processos da comunicação, a estandardizada, àquela que traz clareza e eficiência, mas que no fim é apenas uma das possibilidades de comunicação. (Cabré, 1999 apud Almeida, 2003, p. 214).

De acordo com Cabré et al. (1998, p. 36-37) citado por Almeida (2003, p. 215), a TGT demonstra-se insuficiente nos seguintes aspectos: descrição das relações entre conceitos em um modelo de organização do conhecimento hierárquico e binário; o conhecimento especializado aplica-se da mesma forma à conceitos culturais,

geográficos e entre outros, que são completamente distintos; a TGT não consegue lidar com realidades distintas de domínios e especialidades, sendo assim, uma teoria imaginativa e intangível.

Para Medeiros (1986), na utilização da abordagem à Terminologia Teórica e Aplicada (TTA), pode-se entender como um campo interdisciplinar por meio dos aspectos teóricos e metodológicos. Essa teoria tem a função de utilizar linguagens especializadas na comunicação entre especialistas das diversas áreas do conhecimento, pois a linguagem especializada compartilha características em comum. Como exemplo à essa metodologia têm-se os Sistemas de Organização do Conhecimento (SOC's) que são de um domínio específico, e pode ser de uma área do conhecimento em específico também, ou pode abranger e interligar várias áreas do conhecimento. A pesquisa terminológica no âmbito da TTA compreende as seguintes etapas: coleta, tratamento e difusão de dados terminológicos. Inicialmente, escolhe-se a área do conhecimento que será explorada, escolhe-se o idioma, faz-se a coleta da documentação, levantamento, definição e limitação da quantidade de termos que serão coletados, sempre tendo em vista a objetividade e especificidade desta pesquisa. Essa implementação vem sendo retratada pela autora desde 1986 em uma análise contextual brasileira.

Almeida (2003, p. 216) reúne aspectos sobre a Teoria Comunicativa da Terminologia (TCT) que estão presentes nos trabalhos desenvolvidos por Maria Teresa Cabré entre os anos de 1998 e 1999. A TCT, cuja estrutura teórico-metodológica explica a comunicação especializada e descreve melhor as unidades representativas de forma a desvendar os múltiplos aspectos e camadas intrincadas. Os objetos terminológicos devem ser estudados sob três perspectivas: social, cognitiva e linguística. Sob o aspecto social, leva-se em conta as necessidades comunicativas dos profissionais e dos usuários habitualmente. A perspectiva cognitiva diz respeito ao domínio específico, que é o fenômeno do estudo; e sob a linguística, modelo de atuação e competência.

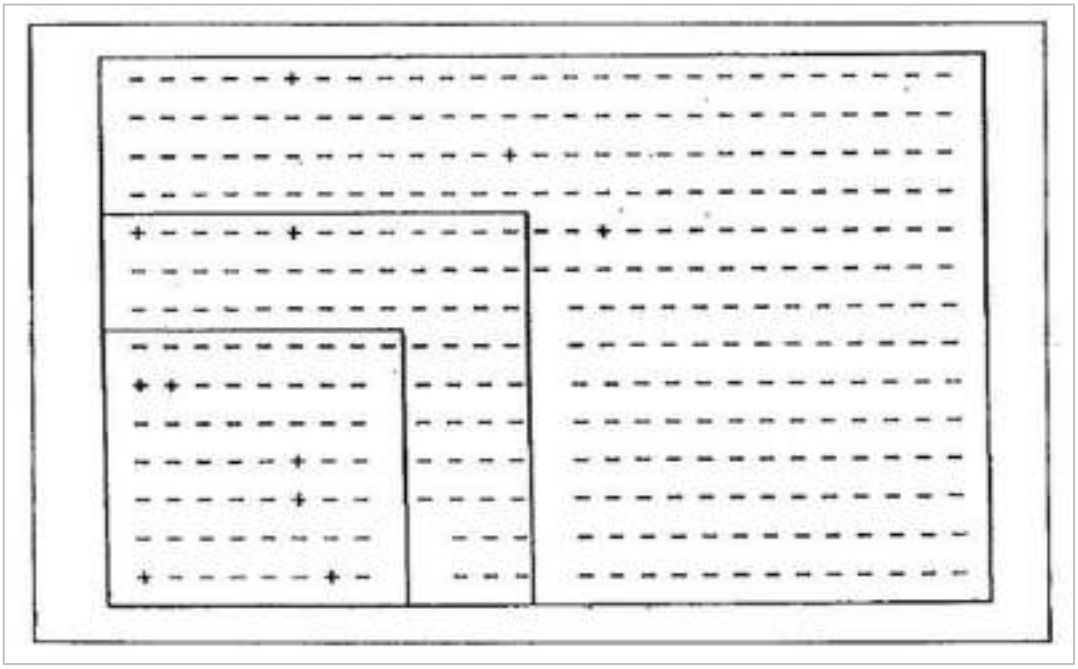
Pereira (2023) em sua monografia de conclusão de curso, extrai por meio automatizado siglas de um corpus documental específico e monta uma lista de termos - vocabulário controlado com menos complexidade do campo terminológico - com conceitos que são definitivamente relacionados ao domínio o qual foi definido. Mediante a essa observação, o termo para ser candidato, também precisa ser capaz de representar o conceito em sua totalidade, devendo ser claro e conciso evitando

ambiguidade, e por consequência, priorizar a conceitualização para a área de conhecimento escolhida.

O trabalho terminológico vai além da melhor escolha dos termos na representação de um documento por meio das metodologias e ferramentas usadas atualmente. O trabalho conhecido como indexação ocorre para que os usuários consigam recuperar um documento de forma eficiente dentro das unidades de informações e bases de dados.

Este trabalho depende do ambiente o qual o profissional está inserido, isto é, se for uma biblioteca universitária por exemplo, ele precisa realizar a indexação com mais proximidade ao preciso possível. Já em bibliotecas especializadas, é necessário que os termos sejam os mais específicos possíveis, assim garante que aquele documento possa ser recuperado com mais facilidade. Em bibliotecas públicas, os profissionais consideram que usar termos mais abrangentes garantem uma indicação maior de documentos aos usuários, isso se dá pelo fato de que em uma biblioteca pública tem diversos tipos de pessoas, de várias idades e de contextos sociais diferentes, o que não é o caso das bibliotecas universitárias e especializadas que tem um nicho menos abrangente de pessoas e interesses, mas ainda assim, é necessário realizar o controle de vocabulário para que os materiais não se percam na imensidão do acervo.

Figura 1 – O problema da recuperação de itens de uma base de dados



Fonte: Lancaster (2004, p. 3)

Lancaster (2004, p. 4) diz que essas técnicas para indexação são necessárias conforme a figura 1, pois dependendo da quantidade de documentos recuperados é possível denominar de duas maneiras: quanto mais preciso for a busca e quanto mais preciso tenha sido indexado o documento, a porcentagem em número de materiais recuperados será menor, e leva o nome de coeficiente de precisão. Quando se recupera uma quantidade abrangente de documentos em uma base de dados, que muitas vezes fora indexado de maneira indevida através de termos que não eram tão pertinentes para representar tal documento, ou que foram usados termos mais genéricos do que precisos, este se denomina como coeficiente de revocação.

Com esses recursos computacionais há vantagem significativa no processo de análise terminológica, consideradas atividades repetitivas, e isso implica regularmente em erros que podem prejudicar o final do processo. Isto é, a automatização programada é capaz de salientar as necessidades do indivíduo que a utiliza, ou seja, na mineração de textos, é possível solicitar ao programa a extração de unidades lexicais especializadas que possam compor futuramente um instrumento de controle terminológico, ou pode também, ser definitivamente uma etapa do trabalho de indexação de objetos informacionais, uma vez que a mineração faz parte um procedimento lógico e metodológico. (Fayyad; Piatetsky-Shapiro; Smyth, 1996, p. 37-51)

Os procedimentos realizados através da mineração de dados são efetivos no quesito de ser rápido, prático, programável, e isso resulta na facilidade da utilização destes sistemas. Para isso usam-se softwares de mineração de textos que fazem com que podemos retirar de dentro de um determinado corpus ou de um determinado documento, aquelas unidades lexicais que se julgam pertinentes para o trabalho terminológico. Ressalta-se que o trabalho terminológico não compreende apenas o âmbito da construção de manutenção de vocabulários controlados, na verdade, fazem parte de um grande sistema que compreende a área da indexação também. São programáveis a partir de algoritmos, realizam a prospecção, agrupagem e outras funções entre os dados, assim revelando novos padrões e relações entre esses termos podendo demonstrar isso de forma visual e gráfica, mesmo não sendo o objetivo deste presente estudo.

> [...] relacionados a um paradigma de programação que nasce da busca por entender as relações textuais, por vezes ambíguas, dos documentos virtuais, que até então não foram resolvidas por outros paradigmas de desenvolvimento de software estando, por sua vez,

> ligadas a diversas outras áreas, dentre elas à Recuperação da Informação. (Medeiros; Pinho; Correa, 2020, p. 154).

Não está ligada apenas à recuperação da informação como atua diretamente aos preceitos da representação da informação, afinal, usando esses métodos de extração de termos, o profissional da CI consegue determinar quais serão aquelas unidades lexicais que representarão um objeto informacional que posteriormente, será objeto da recuperação efetiva da informação.

3. PROCEDIMENTOS METODOLÓGICOS

A metodologia utilizada nesse estudo será a análise de conteúdo, segundo Bardin (1977), com caráter descritivo. Conhecida por muitas significâncias, como análise de assunto, análise temática ou bibliográfica, têm um único propósito: representar a informação e o conhecimento de objetos e documentos informacionais.

A partir dos anos quarenta, inicia-se um evento que dificultou a armazenagem e a recuperação da informação, conhecido pela produção exagerada de conteúdos informacionais. Isso tornou o trabalho do profissional cada vez mais difícil, sendo necessário o desenvolvimento de técnicas adequadas para lidar com o tratamento A metodologia utilizada nesse estudo será a análise de conteúdo, segundo Bardin (1977), com caráter descritivo. Conhecida por muitas significâncias, como análise de assunto, análise temática ou bibliográfica, têm um único propósito: representar a informação e o conhecimento de objetos e documentos informacionais.

A partir dos anos quarenta, inicia-se um evento que dificultou a armazenagem e a recuperação da informação, conhecido pela produção exagerada de conteúdos informacionais. Isso tornou o trabalho do profissional cada vez mais difícil, sendo necessário o desenvolvimento de técnicas adequadas para lidar com o tratamento informacional. Dito isso, o trabalho de representar um documento informacional por meio de termos e conceitos recebe a denominação “indexação”; técnica capaz de extrair os conceitos mediante a leitura e interpretação, e a metodologia mais conhecida e capaz de fazer isso, é a análise de conteúdo, que é considerada uma das análises mais importantes para o trabalho do indexador, realizada por etapas que consistem desde a entrada do material dentro da unidade de informação para que seja determinado o tipo de conteúdo representativo dependendo do ambiente e do usuário, até a recuperação eficaz desse material. (Naves, 1996, p. 215).

No entanto, Bardin (1977, p. 31) diz que independentemente da natureza do suporte ao qual será analisado, a análise de conteúdo tem duas principais funções.

> Uma função heurística: a análise de conteúdo enriquece a tentativa exploratória, aumenta a propensão para a descoberta [...] e uma função de administração da prova. Hipóteses sob a forma de questões ou de afirmações provisórias, servindo de diretrizes, apelarão para o método de análise sistemática para serem verificadas no sentido de uma confirmação ou de uma infirmação. Na prática, as duas funções da análise de conteúdo podem coexistir de maneira complementar. (Bardin, 1977, p. 31-32)

Através de Bardin (1977) observa-se como a análise de conteúdo é uma ferramenta de análise poderosa que pode ser utilizada em diversas áreas do conhecimento, pois com ela conseguimos associá-la em quais meios que se aplica e suas relações com outras ciências, portanto, Bardin define a análise de conteúdo:

> [...] é um conjunto de técnicas de análise das comunicações. Não se trata de um instrumento, mas de um leque de apetrechos; ou com maior rigor, será um único instrumento, mas marcado por uma grande disparidade de formas e adaptável a um campo de aplicação muito vasto: as comunicações. [...] em última análise, qualquer comunicação, isto é, qualquer veículo de significados de um emissor para um receptor controlado ou não por este, deveria poder ser escrito, decifrado pelas técnicas de análise de conteúdo. (Bardin, 1977. p. 33-34).

Para Albrechtsen (1993) existem três pontos principais que caracterizam a análise de conteúdo:

a) Concepção Simplista – onde a partir da análise textual é possível a extração de unidades lexicais que representam o conteúdo;
b) Concepção Orientada para o Conteúdo – é possível identificar conceitos que representam o tema a partir da leitura técnica, mas que esses conceitos não necessariamente estão presentes dentro do documento. Assim, o profissional determina a partir do ambiente informacional e de seus usuários qual a melhor forma de representar aquele documento;
c) Concepção Orientada pela Demanda – aqui, o indexador utiliza de uma técnica que faz uma representação a partir da mediação do assunto de modo que o usuário se interesse por aquele material.

Esses três pontos são considerados essenciais para o trabalho do indexador porque são complementares, mesmo que a concepção orientada pela demanda seja uma pós etapa daquelas consideradas iniciais para este trabalho.

Este presente estudo buscar utilizar da análise de conteúdo com caráter descritivo analítico, que para Bardin (1977, p. 37) “A descrição analítica funciona segundo procedimentos sistemáticos e objetivos da descrição do conteúdo das mensagens. Trata-se, portanto, de um tratamento da informação. ” A análise teórica será por meio de artigos científicos, com propósito de aprofundar-se na qualidade da particularidade que pode ser abstraído de cada um destes softwares, que estes serão o Sistema de Indización Semi-Automático (SISA) e o software de extração de termos em objetos textuais, o Sodek.

Para realizar este estudo comparativo é necessário seguir uma sequência de etapas que se constituem em:

a) Determinar os objetivos do trabalho e o escopo, reunindo os aspectos teóricos, selecionar os documentos que serão analisados e realçar as características;
b) Coletar e preparar os dados através de uma organização prévia destes documentos e sistematizá-los ou categorizá-los para facilitar a análise;
c) Interpretar os dados através de suas características, apontando semelhanças e diferenças, pontos positivos e negativos;
d) Redigir detalhadamente por meio de um código suporte e delimitando como a comunicação será feita (baseado na tabela *“Domínios possíveis da aplicação da análise de conteúdo”* segundo Bardin, 1977, p. 36);
e) Apresentar os resultados.

4. RESULTADOS: APRESENTAÇÃO E DISCUSSÃO

Os resultados apresentam a potencialidade dos softwares escolhidos na prospecção de unidades lexicais que realizam a representação e recuperação de objetos e documentos informacionais, e softwares que realizam a indexação automática (ou pelo menos parte dela), e em quais os casos é necessária a ação humana para interferir nesse processo.

4. 1SISTEMA DE INDIZACIÓN SEMI-AUTOMÁTICO (SISA) – SOFTWARE DE INDEXAÇÃO

Criado por Gil Leiva em 1999 a 2008 na Espanha com propósito de ser um sistema de indexação semiautomatizado capaz de ler o conteúdo principal de um documento, tal como título, resumo e texto, concebido inicialmente para as áreas da

biblioteconomia e documentação. Contudo, é um sistema totalmente flexível a adaptar sua aplicação em outras áreas do conhecimento. (Narukawa; Gil Leiva; Fujita, 2009.p. 106).

Para os autores, os segmentos do procedimento do SISA compreendem-se em:

a) O pré-processamento, onde o sistema indica as partes com marcadores, onde inicia-se o título, resumo ou texto, e onde termina, também faz a limpeza no campo das palavras vazias;

b) A análise do conteúdo, onde através de um algoritmo é possível buscar por termos preferidos a apresentarem o documento. Dessa forma é possível fazer a exclusão daqueles que não serão pertencentes, tal como termos ambíguos, sinônimos, ou aqueles que o próprio sistema julga como não pertinente à representação;

c) Nesta etapa, selecionam-se os termos autorizados a apresentarem o documento. O próprio sistema faz essa escolha mediante a algumas regras, tais como termos que apareçam na fonte do título, do resumo ou do texto. Isto é, se um documento tem o título de *"Metafísica do Poder",* os termos considerados pertinentes serão *"metafísica"* e *"poder",* pois aparecerão em ambas as partes do documento em uma quantidade de vezes exageradas;

Já nesta etapa, a indexação é caracterizada como semiautomatizada pois depende da interferência humana como decisão do profissional, ou seja, aqui o terminológo/indexador tem o poder de determinar quais outros termos são pertinentes e quais aqueles que não serão necessários para a representação.

Os benefícios encontrados através da leitura sobre o software SISA, segmentam-se na produtividade do profissional da informação, ou seja, este software permite que o profissional tenha maior desempenho em suas atividades terminológicas a partir do princípio de que a máquina é capaz de ser mais rápida na leitura documental, conseguindo de forma abrangente e exaustiva um número de unidades lexicais especializadas que possam fazer parte da representação do documento. Não como um ponto negativo, mas que reforça a competência profissional, é que por mais que a máquina trabalhe realizando as principais etapas da extração até a indexação dos termos (duas áreas em conjunção), é de extrema importância ressaltar que a capacidade humana de aplicar conhecimento sobre esses procedimentos é imprescindível, tendo em vista que humanos têm percepção mais ampla e um conhecimento mais significativo sendo capazes de determinar de fato o

que é ou o que não é pertinente à representar tal material.

4.2. SOBEK – SOFTWARE DE MINERAÇÃO DE TEXTOS

Este software de mineração foi desenvolvido pelo programa de pós-graduação em Informática na Educação pela Universidade Federal do Rio Grande do Sul (UFRGS). Segundo o site da Universidade Federal do Rio Grande do Sul (2007) este software é capaz de "identificar os conceitos relevantes em um texto a partir da análise de frequência no material" através da mineração de dados que se torna popular devido ao crescimento exponencial de informação na internet contemporânea. Criado em 2007 com o propósito de ser uma ferramenta de mineração que auxiliasse os professores no ensino à distância. Em 2009, começou a ser usado por alunos para a compreensão de leitura e a realização de resumos. Em 2010, foi incorporado em outros sistemas com a proposta de melhorar a narrativa escrita, como uma ferramenta de aprendizagem.

> Frente a uma sempre crescente quantidade de material escrito por alunos e disponíveis para pesquisa, além de um aumento expressivo no ensino a distância (EaD) nos últimos anos, percebeu-se que havia a necessidade de auxiliar professores que ministravam cursos EaD a classificarem e analisarem textos de maneira mais rápida e automática. (Epstein, 2017, p. 49).

Por ser um software de mineração de dados, nos propõe a busca de padrões dentro dos textos e em banco de dados não estruturados. Ele funciona como uma ferramenta nesse sentido, porque a mineração de dados, na verdade, é uma etapa de um processo maior, mas cabe a cada profissional delimitar sua funcionalidade.

Um ponto positivo do Sobek, é que este software após realizar a mineração do texto, faz em representação manual através de um grafo a frequência dos termos, ou seja, os nós que estão maiores, representam a maior quantidade dentro de um texto. Nessa representação gráfica, também é possível visualizar como os termos estão relacionados entre si. Outro ponto positivo, exemplificado dentro de seu manual, é que através deste software de mineração, consegue-se a criação de Tesauros.

Para Epstein (2017, p.49) o Sobek "possui como principal característica uma interface simples, pensada para permitir o seu uso sem a necessidade de treinamento prévio ou de conhecimentos de informática", e isso implica na facilidade de usabilidade do software, como qualquer pessoa possa utilizar para qualquer finalidade no âmbito

da mineração de textos.

Para atingir as proposições com o suporte da análise de conteúdo de Bardin (1977), fora elaborado um questionário (com caráter icônico, por meio da comunicação dual) sobre as características dos softwares, onde através das respostas de cada uma das perguntas será possível realizar análise comparativa entre suas similaridades e diferenças, permitindo o apontamento dos benefícios e malefícios encontrados neles através da análise teórica.

Quadro 1 - Comparativo entre as características dos softwares SISA e Sobek

	SISA	**SOBEK**
O que é este Software	Ferramenta de indexação semiautomatizada	Ferramenta de mineração de textos
Para qual finalidade foi criada?	Criado para analisar textos em português por meio da extração de sintagmas nominais e cálculo do peso desses na indexação dos documentos. (Silva;Correa, 2020, p. 9)	Criada para auxiliar professores no ensino a distância, que viabilizasse em menor tempo o gerenciamento de um alto volume de dados gerados nas produções textuais. (Epstein, 2017, p. 49)
Qual a qualidade de extração de informações dos softwares SISA e Sobek?	Os requisitos de entrada no SISA são: lista alfabética de termos e descritores e respectivos termos gerais, lista de palavras vazias no idioma do texto dos documentos para fins de eliminação de stop words. (Silva; Correa, 2020, p. 12-13)	Com base nas necessidades, foram implementadas configurações de extração de termos: seleção de frequência mínima, número médio de termos, lista de stop words e suporte aos idiomas português e inglês. (Epstein, 2017, p. 51)
Qual a capacidade de análise de textos deles?	Todos os arquivos de entrada, incluindo os textos a serem indexados, devem estar em formato *.txt.* (Silva; Correa, 2020, p. 13)	Além da extração de termos do texto copiados para a caixa de edição, é possível abrir um texto salvo no computador ou mesmo uma mineração prévia. Suporta os formatos *.doc e .pdf.* Também permite a análise de um conjunto de textos em diferentes formatos. (Epstein, 2017, p. 51)
Qual a facilidade de uso?	Interface pouco intuitiva, usado por profissionais terminólogos que tem foco na indexação automatizada a partir da extração de termos de um texto. (Narukawa; Gil Leiva; Fujita, 2009, p. 105-106)	Interface simplificada, permitindo que qualquer profissional seja ele professor ou terminólogo consiga utilizar sem muita burocracia. (Epstein, 2017, p. 50)
Existe flexibilidade na adaptação destes softwares?	Limitado, pois seu foco é a indexação de termos, mas é extensível para outros idiomas. (Silva; Correa, 2020, p. 18)	Abrange flexibilidade permitindo ser usado em diversos tipos de projetos e necessidades à mineração de textos.

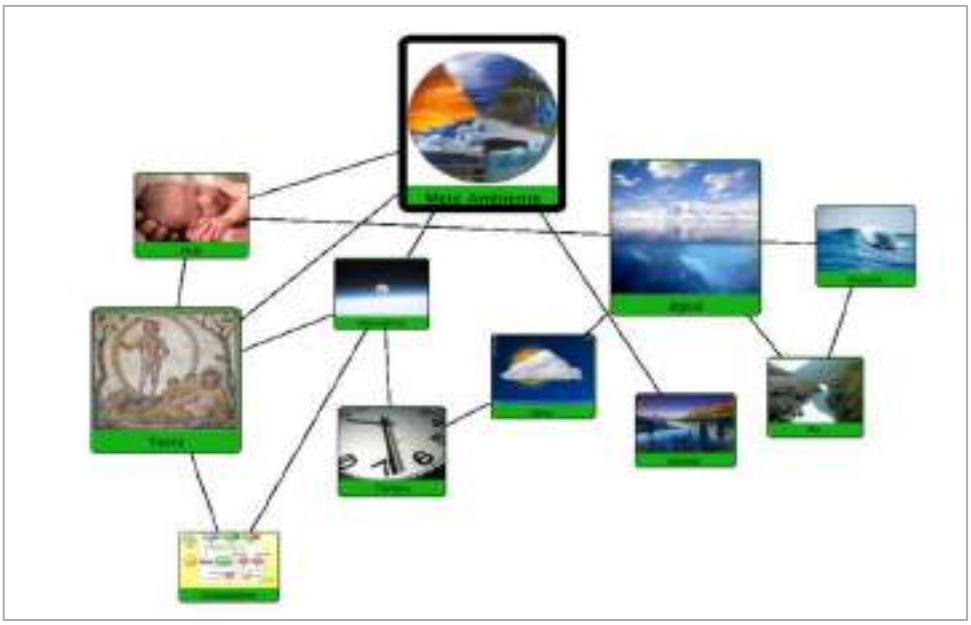
Fonte: Elaborado pelos autores, 2024

Através desta comparação, considera-se que ambos os softwares possuem pontos positivos e pontos negativos. Os pontos positivos do SISA é justamente sua

precisão bem estruturada sendo um programa ideal para indexar documentos, muito útil para bibliotecas e bases de dados. Os pontos negativos se dão pela falta de compatibilidade com outros formatos de documentos, limitado quanto a capacidade de análise de textos, e interface pouco intuitiva.

Para o Sobek, o ponto positivo é ser uma ferramenta de livre acesso, sendo possível qualquer pessoa podendo usá-la para diferentes finalidades, não sendo totalmente voltado para a área da biblioteconomia ou documentação, tanto que foi idealizado para a área da educação. Como o software Sobek tem a proposta de auxiliar profissionais e alunos em diversos tipos de projeto, Reategui, Campelo e Oliveira (2017) analisam o Sobek em um estudo que tem como objetivo investigar a ferramenta de mineração de texto para qualificar os textos acadêmicos. Os autores consideram que a ferramenta SOBEK como mecanismo de Diagnóstico de Aprendizagem alcança o seu objetivo principal, do qual faz com que o aluno reflita sobre sua forma de escrita a partir dos grafos explorando a autonomia do próprio aluno. Uma das dificuldades ressaltadas pelos autores está na utilização ou do download da ferramenta, erros que podem ser corrigidos posteriormente. Costa *et al.* (2017) desenvolveram um estudo com objetivo de analisar se o SOBEK poderia contribuir no processo de construção de conceitos científicos. O benefício do software citado por Costa *et al* (2017, p. 94) está na pós mineração do texto, onde o “software realiza uma busca na internet por imagens associadas a cada um dos termos selecionados e as apresenta no grafo, junto ao termo correspondente” conforme a figura a seguir.

Figura 2 – Grafo gerado com a ferramenta SOBEK a partir de um texto selecionado pelos autores



Fonte: Costa *et al* (2017, p. 95)

Costa *et al* (2017, p. 105) constatam que “a utilização do software aponta a importância da experimentação da inserção tecnológica e de abordagens inovadoras nas práticas educacionais” e para além, concluem que “[...] o trabalho de interpretação e edição do grafo poderia estimular a interação do aprendiz com o seu material de estudo, pois nesse processo, o texto escrito vai sendo explorado, reinterpretado e representado de formas diferentes”. No entanto, o estudo destes autores aponta uma falha nos testes empregados, onde afirmam não conseguirem captar aprimoramentos das habilidades dos participantes em novas situações.

Os softwares SISA e Sobek são boas ferramentas para o segmento da indexação e mineração de termos com finalidade de recuperação de documentos informacionais. O SISA permite uma eficiente indexação semiautomatizada de documentos, facilitando a organização e a busca de informações relevantes. Já o Sobek, além de sua capacidade de mineração de textos e representação visual em formato de grafo, destaca-se em outros pontos, como o suporte a professores na correção de atividades no contexto do ensino a distância (EAD), essa forma, ambos os softwares proporcionam benefícios significativos.

5. CONSIDERAÇÕES FINAIS

Mediante aos resultados apresentados, conclui-se que a prática terminológica, em especial o da indexação, não é totalmente automatizada, pelo fator de que as tecnologias não produzem conhecimento, apenas nos dão dados e informações que

já estão pré-estabelecidas em bancos de dados ou quando são programadas para executar funções, afinal, a aplicação de conhecimento é proveniente da competência profissional. No entanto, são ferramentas essenciais na atualidade, já que através delas o profissional consegue ganhar tempo ao lidar com uma alta demanda de trabalho e nas tarefas que se considera repetitivas e demoradas, como a análise técnica dos documentos. Haja vista que muitas vezes, em uma leitura técnica do documento, pode não ser suficiente para extrair unidades lexicais especializadas à representar o objeto informacional, e que pode gerar vício de leitura do profissional.

Observa-se que o software SISA contribui grandemente com o processo de indexação semiautomatizada das unidades lexicais extraídas por meio de análise de frequência dentro do texto, possibilitando a facilidade da representação e armazenação documentária, e a posteriori, a classificação e recuperação do documento de forma eficaz.

O software SOBEK atuaria muito bem como ferramenta auxiliadora na construção de vocabulários controlados. Este estudo comparativo não se aprofunda nas questões de construção de vocabulário controlado, e futuramente uma análise mais aprofundada do desempenho destes programas poderá ser realizada. A priori, verificou-se que mediante a análise teórica sobre os softwares, que o minerados de textos SOBEK pode estar presente em diversos tipos de projetos. Por realizar uma rede entre os termos principais encontrados dentro do documento, além de ajudar na construção de vocabulário controlado, pode ser realizado um estudo da relação entre os termos dentro de um domínio específico, pois nos garante um código de suporte icônico (representação gráfica) por meio de um grafo.

Contudo, confirma-se que a utilização das tecnologias beneficia as práticas terminológicas pois conseguem manipular uma quantidade exorbitante de dados, no âmbito que a informação cresce exponencialmente a cada minuto. Porém, é imprescindível o trabalho terminológico na esfera de sua competência, pois traz garantia literária e precisão destas práticas, facilitando aos usuários uma busca e recuperação eficiente dos resultados esperados.

Ainda que a mineração, seja de dados ou textual, ela está inserida como uma etapa de um processo maior: a análise de dados, tendo em vista que o termo prospectado só ganha certificação de “informação” quando passa pelo processo de tratamento da informação mediante à metodologia, que neste caso, utilizou da análise de conteúdo proposta por Laurence Bardin (1977). A análise de conteúdo está

presente no processo de análise teórica na demonstração dos resultados, e no processo de mineração textual que os softwares apresentados fazem, pois, a análise de conteúdo é responsável por trazer sentido e contexto à informação por meio de suas características.

Por fim, esta pesquisa proporciona uma base sólida para a escolha do software que se adequa às necessidades das práticas terminológicas.

REFERÊNCIAS

ALBRECHTSEN, Hanne. Subject Analysis and Indexing: from automated indexing to domain analysis. **The Indexer,** Londres, v. 18, n. 4, p. 219-224, 1993. Disponível em: https://www.researchgate.net/publication/265191594_Subject_analysis_and_indexing-From_automated_indexing_to_domain_analysis#fullTextFileContent. Acesso em: 4 jul. 2024.

ALMEIDA, Gladis Maria de Barcellos. O percurso da Terminologia: de atividade prática à consolidação de uma disciplina autônoma. **Tradterm**, São Paulo, v. 9, p. 211-222, 2003. Disponível em: https://revistas.usp.br/tradterm/article/view/49087. Acesso em: 10 jun. 2024.

ALMEIDA, Gladis Maria de Barcellos; OLIVEIRA, Leandro Henrique Mendonça de; ALUÍSIO, Sandra Maria. A terminologia na era da informática. **Ciência e Cultura**: São Paulo, v. 2, p. 42-45, junho de 2006. Disponível em http://cienciaecultura.bvs.br/scielo.php?script=sci_arttext&pid=S0009-67252006000200016&lng=en&nrm=iso . Acesso em 23 de maio de 2024.

BARDIN, Laurence. **Análise de Conteúdo**: concepções. Edições 70 Lda: Lisboa, 4. ed, p. 281, 1977.

CABRÉ, Maria Teresa. **La Terminología – representación y comunicación**: elementos para uma teoria de base comunicativa y otros artículos. Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada (IULA), 1999. Disponível em: https://dialnet.unirioja.es/servlet/articulo?codigo=6468207. Acesso em: 10 jun. 2024.

COSTA, Ana Paula Metz; REATEGUI, Eliseo Berni; EPSTEIN, Daniel; MEYER, Daniel Derrossi; LIMA, Evelyn Gonçalves; SILVA, Karina Heck da. Emprego de um software baseado em mineração de texto e apresentação gráfica

multirrepresentacional como apoio à aprendizagem de conceitos científicos a partir de textos no Ensino Fundamental. **Ciência & Educação,** Bauru, v. 23, n. 1, p. 91-109, 2017. Disponível em: https://www.scielo.br/j/ciedu/a/nwKvB7d533sLx5Pjdhfg6JJ/?format=pdf. Acesso em: 3 jul. 2024.

EPSTEIN, Daniel. **Uso do minerador de textos Sobek como ferramenta de apoio à compreensão textual**. 2017. Tese (Doutorado) – Universidade Federal do Rio Grande do Sul, Centro de Estudos Interdisciplinares em Novas Tecnologias em Informática na Educação, Porto Alegre, 2017. Disponível em: https://lume.ufrgs.br/bitstream/handle/10183/178332/001066224.pdf?sequence=1&isAllowed=y. Acesso em: 08 jun. 2024.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge Discovery in Databases. **AI Magazine**, v. 17, n.3, 1996. Disponível em: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230. Acesso em: 10 jun. 2024.

KRIEGER, Maria da Graça. **Terminologias em Construção**: procedimento metodológicos. TERMSUL-UFRGS, 2005. Disponível em: https://www.ufrgs.br/termisul/files/file112160.pdf. Acesso em: 10 jun. 2024.

LANCASTER, Frederick Wilfrid. **Indexação e resumos**: teoria e prática. 2. ed. Brasília: Brinquet Lemos, 2004. Disponível em: https://bibliotextos.wordpress.com/wp-content/uploads/2014/07/livro-indexac3a7c3a3o-e-resumos-teoria-e-prc3a1tica-lancaster.pdf. Acesso em: 5 jun. 2024.

MEDEIROS, Marisa Bräscher Basílio. Terminologia Brasileira em Ciência da Informação: uma análise. **Ciência da Informação**, Brasília, v. 15, n. 2, 1986. Disponível em: https://revista.ibict.br/ciinf/article/view/234. Acesso em 10 jun. 2024.

MEDEIROS, Wagner Oliveira de; PINHO, Fabio Assis; CORREA, Renato Fernandes. Aplicação de software de mineração de texto na representação da informação de obras artístico-pictóricas. Logeion: **filosofia da informação**, v. 6, n. 1, 2019. Disponível em: https://www.brapci.inf.br/#/v/121967. Acesso em: 07 jun. 2024.

NARUKAWA, Cristina Miyuki; GIL LEIVA, Isidoro Gil; FUJITA, Mariângela Spotti Lopes. Indexação automatizada de artigos de periódicos científicos: análise da aplicação do software SISA com uso da terminologia DeCS na área de Odontologia. **Informação & Sociedade**, Londrina, v. 19, n. 2, 2009. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/2925. Acesso em: 10 jun. 2024.

NAVES, Madalena Martins Lopes. Análise de Assunto: concepções. **R. Biblioteconomia**, Brasília, v. 20, n. 2, p. 215-226, 1996.

PEREIRA, L. K. P. B. P. P. **Prospecção de Termos Candidatos no Domínio do Transtorno do Espectro Autista (TEA).** 2023. 69 f. TCC (Bacharel em

Biblioteconomia) – Centro de Educação, Comunicação e Artes, Universidade Estadual de Londrina, Londrina, 2023.

REATEGUI, Eliseo Berni; CAMPELO, Patrícia; OLIVEIRA, Simone de. O Apoio de uma ferramenta com base na mineração de texto para escrita acadêmica. **Informática na Educação: teoria e prática,** Porto Alegre, v. 20, n. 1, 2017. Disponível em: https://seer.ufrgs.br/index.php/InfEducTeoriaPratica/article/view/70063/41071. Acesso em: 3 jul. 2024.

SILVA, Sâmela Rouse de Brito; CORREA, Renato Fernandes. Sistemas de Indexação automática por atribuição: uma análise comparativa. **Encontros Bibli:** revista eletrônica de biblioteconomia e ciência da informação, v. 25, p. 01–25, 2020. DOI: 10.5007/1518-2924.2020.e70740. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2020.e70740. Acesso em: 10 jun. 2024.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL. **Sobek Mining**. 2010. Disponível em: http://sobek.ufrgs.br/#/about. Acesso em: 24 jul. 2024.