Ciência da Informação em Revista
ISSN: 2358-0763

O comportamento do termo Dado na Ciência da Informação

The behavior of the term Data in Information Science

Marcos de Souza
Doutor em Gestão e Organização do Conhecimento
Universidade Federal de Minas Gerais
marcosdesouza82@gmail.com

Fernanda Gomes Almeida
Doutora em Gestão e Organização do Conhecimento
Universidade Federal de Minas Gerais
usernanda@gmail.com

Resumo
Advindos de diferentes áreas, os termos dado, informação e conhecimento se tornaram objeto de estudo da Ciência da Informação, principalmente após o surgimento das Tecnologias da Informação. O termo dado, isoladamente, apresenta significado restrito e pouco informativo, sendo puramente objetivo, porém, representa matéria-prima para uma série de observações, medidas ou fatos. A pesquisa buscou identificar, por meio de frequências de comportamento, o comportamento do termo dado na área da Ciência da Informação, ao longo do período estudado e os seus principais termos formados por composições, além de criar uma classificação para os tipos de comportamentos. Foram realizadas, durante a pesquisa empírica, as etapas de coleta de dados, preparação e pré-processamento, transformação e apresentação dos dados, que resultou em uma frequência de termos extraídos do *corpus*. Considera-se, dentre os resultados, que o termo dado se apresenta, através de seu comportamento, de forma contínua e em ascensão ao longo do período analisado, entretanto, os termos compostos apresentam uma maior especificidade de domínio da linguagem, quando se comparado ao termo raiz como, por exemplo, os termos “base_dados”, “dados_pesquisa”, “coleta_dados”, “fonte_dados” e “banco_dados”, que somados, ultrapassam mais de 50% do termo raiz. O termo “gestão_dados” apresentou comportamento irregular em ascensão, o que permite inferir que pesquisas que abordam a temática começaram a ganhar notoriedade na área da Ciência da Informação nos últimos analisados.

Palavra-chave
Dado. Comportamento. Termo. Ciência da Informação.

Abstract
Coming from different areas, the terms data, information and knowledge have become the object of study of Information Science, especially after the emergence of Information Technology. The term data, in isolation, presents a restricted and uninformative meaning, being purely objective, however, it represents raw material for a series of observations, measures or facts. The research sought to identify, by means of behavioral frequencies, the behavior of the term data in the area of Information Science, throughout the period studied and its main terms formed by compositions, in addition to creating a classification for the types of behavior. During the empirical research, the steps of data collection, preparation and pre-processing, transformation and presentation of the data were carried out, which resulted in a frequency of terms extracted from the corpus. The results show that the term data presents itself, through its behavior, in a continuous and rising form throughout the analyzed period; however, the compound terms present a greater specificity of language domain when compared to the root term, such as, for example, the terms "base_dados", “dados_pesquisa”,

DOI: 10.28998/cirev.2021v8n2c

Submetido em: 24/01/2021
Aceito em: 03/11/2021
Publicado em: 12/11/2021

"coleta_dados", "fonte_dados" e "banco_dados", which added together, exceed more than 50% of the root term. The term "gestão_dados" showed irregular behavior on the rise, which allows us to infer that research addressing the topic has begun to gain notoriety in the area of Information Science in the last analyzed.

Keyword
Data. Behavior. Term. Information Science.

1 INTRODUÇÃO

Os conceitos de dado, informação e conhecimento têm sido discutidos ao longo dos anos por diferentes autores nos aspectos epistemológicos, etimológicos, filosóficos e de aplicações em áreas como Administração, Economia, Políticas Sociais e Tecnologias da Informação e Comunicação.

Os termos informação e conhecimento apresentam caminhos diferentes ao longo da história, sendo o primeiro, advindo das Ciências Exatas, enquanto conceito matemático, para tratar a comunicação e, o segundo, surgido da Filosofia e Sociologia. Ambos os termos são associados às tecnologias da informação e se tornaram objetos de estudos da Ciência da Informação juntamente com o termo dado, que passa a ser estudado nas Ciências Sociais (SIRIHAL; LOURENÇO, 2002).

A diferenciação entre os conceitos de dado, informação e conhecimento pode se tornar uma tarefa difícil, justamente por não possuir uma convergência de ideias entre autores em torno dos pontos que delimitam - começam e terminam - a representatividade dos conceitos. Enfatiza-se que, a não uniformidade entre os conceitos quando aplicado a Ciência da Informação se dá, justamente, por ser uma área relativamente nova e interdisciplinar, que recebe diferentes contribuições em seu arcabouço teórico. Apesar de ainda não haver uma convergência entre os autores, a tríade, enquanto processo, permite considerar que os dados podem ser utilizados em diferentes situações, as hipóteses são utilizadas para estruturação da informação e, finalmente, as tomadas de decisões são realizadas com base no conhecimento (DAVENPORT, 1998).

Levando em consideração pesquisas científicas já publicadas que abordam os conceitos de dado, informação e conhecimento, este artigo optou por explorar o comportamento do termo dado ao longo do tempo, na área de Ciência da Informação. Partindo deste princípio, questiona-se: de que forma tem se apresentado o comportamento do termo dado na Ciência da Informação? O objetivo da pesquisa é: 1) identificar o comportamento do termo dado bem como suas principais composições de *n*-grama, formadas a partir do termo analisado e seus respectivos comportamentos e; 2) criar uma classificação de comportamento para os termos.

Entende-se por composição a concatenação de, ao menos, duas bases que podem ser autônomas e supostamente com capacidade referencial, sendo o termo composto, uma unidade lexical formada a partir de duas unidades lexicais dotadas dos seus respectivos referenciais (RIO-TORTO, 1998). São exemplos de composição os termos "banco de dados" e "base de dados". Já o *n*-grama consiste em um pedaço de *n*-caracteres, extraídos de uma cadeia de caracteres. Para *n*, assume-se valores como 1, 2 ou 3, respectivamente para unigrama, bigrama e trigrama (SUKKARIEH; PULMAN; RAIKES, 2003).

Para Paschoalin e Spadoto (1996) a linguagem natural é variável, com palavras que caem em desuso, mudam de significados ou surgem com o passar dos anos. Assim, o vocabulário de uma área em constante evolução.

Pressupõe-se que o termo "dado", analisado em um *corpus*, possa apresentar comportamentos que não remetem a realidade de sua representatividade enquanto conceito, uma vez que, composições de palavras são realizadas para a formação de outros termos e conceitos diferentes ao de origem. Dessa forma, justifica-se a importância da pesquisa, uma vez que o mapeamento científico do termo pode apontar composições de termos com outros significados indicando, a ascensão ou descensão a partir do termo raiz em uma linguagem de domínio.

2 REFERENCIAL TEÓRICO

A emergência dos termos que compõem a tríade dado, informação e conhecimento na Ciência da Informação iniciada a partir de Borko (1968), as origens de cada elemento da tríade apresentado por Sirihal e Lourenço (2002), bem como as características interdisciplinares da Ciência da Informação apresentadas por Saracevic (1996) são norte para o estudo do comportamento do termo "dado" apresentado neste trabalho.

Na concepção do texto *Information science: what is it?* de Borko (1968) encontra-se uma das primeiras e mais reconhecidas definições do termo Ciência da Informação, que podem ser apresentadas em três estruturas, mediante sua densidade e pluralidade.

A primeira estrutura afirma que a Ciência da Informação, enquanto disciplina, investiga tanto as propriedades quanto o comportamento da informação, bem como seu fluxo de governança e os meios de processos e otimização de sua acessibilidade e uso. A segunda preocupa-se com o corpo do conhecimento relacionado a um conjunto de procedimentos que envolve a informação sendo: organização, armazenamento, recuperação, interpretação, transmissão, transformação e utilização. Já a terceira estrutura apresenta componentes da ciência pura, sem a necessidade de atentar para uma aplicação e componentes da ciência aplicada, ao qual busca desenvolver produtos e serviços (BORKO, 1968).

Em seu processo evolutivo, a Ciência da Informação apresenta três características para a compreensão do passado, presente e futuro da área, sendo a primeira relacionada à interdisciplinaridade, em que, mesmo havendo mudanças entre as relações de disciplinas, não existe perspectivas para a evolução interdisciplinar se completar. A segunda está intrinsecamente associada à tecnologia da informação, que reflete a transformação da sociedade moderna para sociedade da informação. A terceira está relacionada a tantas outras disciplinas envolvidas na evolução da sociedade da informação (SARACEVIC 1996).

Saracevic (1996) enfatiza que a Ciência da Informação apresenta um importante papel nas dimensões sociais e humanas que ultrapassam as tecnologias ao conceituar o termo.

> A CIÊNCIA DA INFORMAÇÃO é um campo dedicado às questões científicas e à prática profissional voltadas para os problemas da efetiva comunicação do conhecimento e de seus registros entre os seres humanos, no contexto social, institucional ou individual do uso e das necessidades de informação. No tratamento destas questões são consideradas de particular interesse as vantagens das modernas tecnologias informacionais (SARACEVIC,1996, p. 47).

A interdisciplinaridade na Ciência da Informação se deu a partir das diferentes experiências de áreas que buscavam/buscam por resolutivas para seus respectivos problemas. Embora as diferentes experiências apresentem riquezas no campo estudado, nem todas as disciplinas apresentam contribuições relevantes, mas tal variedade é responsável por caracterizar a interdisciplinaridade na Ciência da Informação (SARACEVIC, 1995).

A Ciência da Informação apresenta duas vertentes, sendo a primeira na bibliografia/documentação, onde o foco está no registro do conhecimento científico, na memória intelectual da civilização e, a segunda, na recuperação da informação, realizada por meio de aplicações tecnológicas em sistemas de informação. A Ciência e a Tecnologia foram elementos primordiais para o nascimento da área, sobretudo a partir da Segunda Guerra Mundial (PINHEIRO, 2005).

A emergência dos termos que compõe a tríade dado, informação e conhecimento na Ciência da Informação poderia ser reconhecida historicamente a partir dos esforços de fundamentação disciplinar responsáveis pela elaboração identitária da Ciência da Informação, realizada por meio das contribuições de Borko (1968), que delineou requisitos científicos da nova disciplina embasado na demanda por resoluções de problemas de informação e conhecimento (SEMIDÃO, 2013).

A diferenciação entre dado, informação e conhecimento é um processo de difícil definição, entretanto, elaborar um processo que relaciona os três conceitos permite a busca por resultados diferenciados (DAVENPORT, 1998). No contexto da Ciência da Informação e embasados por autores da área, Pinheiro (2005) destaca que, ao final do processo da tríade dado, informação e conhecimento, poderá gerar uma fase final, que é a cultura do ser homem, seja construída individualmente ou coletivamente:

> A cadeia conceitual que caracteriza a Ciência da Informação vai desde o dado à informação e conhecimento, de acordo com a ideia de muitos de seus autores, algumas vezes incluindo saber; num crescendo de complexidade, da forma bruta e primitiva do dado à sua elaboração como informação, e sua absorção, quando relevante, na estrutura cognitiva, transformando-se em conhecimento. Esta rede de conceitos poderá ter seu processo final na cultura, aqui considerando a incorporação dessas informações relevantes entre outras manifestações e produções e vivências do homem, individuais e coletivas. (PINHEIRO, 2005, p. 40).

O termo conhecimento, abordado inicialmente pelas ciências humanas, sobretudo na filosofia e sociologia, juntamente com o termo informação, abordado nas ciências exatas no contexto da teoria matemática da comunicação (SIRIHAL; LOURENÇO, 2002) ganharam relevância socioeconômica sob uma nova ótica (SEMIDÃO, 2014), tornando-se recurso econômico fundamental. Informação e conhecimento passaram a ser de interesse também pela Ciência da Informação. Sendo assim:

> [...] o conceito de informação foi trazido para as ciências sociais e começou a ser trabalhado juntamente com o conceito de conhecimento. Neste universo, surge o termo dado, que passa a ser objeto de estudo também das Ciências Sociais (SIRIHAL; LOURENÇO, 2002, p. 1-2).

Os dados não apresentam significados próprios, porém, representam a matéria-prima como uma série de observações, medidas ou fatos que podem ser representados como números, palavras, sons ou imagens ao qual é produzida a informação. Para possuir um significado, a informação deve estar relacionada em um determinado contexto, sendo os dados organizados de maneira significativa. Já o conhecimento é definido como a aplicação e o uso produtivo da informação (BOISOT, 1998).

Dentre as características dos termos da tríade, considera-se o dado puramente objetivo, o que independe do usuário. A informação objetiva-subjetiva, sendo descrita de uma forma objetiva com significados subjetivos. O conhecimento é puramente subjetivo, considerando a vivência do indivíduo (SETZER, 1999).

O *Oxford Advanced Learner's Dictionary of Current English* define dados como: "1 fatos ou informações, especialmente quando examinados e usados para descobrir coisas ou para tomar decisões [...]. 2 informações que são armazenadas por um computador [...]" (DATA, 2005, p. 387, tradução nossa). Já o *Online Dictionary for Library and Information Science* define dados como:

> O plural da palavra latina *datum* que significa "o que é dado", frequentemente usada como um substantivo coletivo singular. Fatos, números ou instruções apresentadas de uma forma que possam ser compreendidas, interpretadas e comunicadas por um ser humano ou processadas por um computador (DATA, 2020, on-line, tradução nossa).

Cunha e Cavalcanti (2008, p.112-113) consideram o termo dado como "a menor representação convencional e fundamental de uma informação (fato, noção, objeto, nome próprio, número, estatística, etc.) sob forma analógica ou digital" podendo ainda ser submetido a processamentos manuais ou automáticos.

Para Cunha e Cavalcanti (2008) o termo dados se caracteriza em um sentido mais amplo como:

> [...] toda informação quantificável (números, letras, gráficos, imagens, sons ou uma combinação desses tipos). 1.1 Sinais ou códigos usados para alimentação, armazenamento, processamento e produção de um resultado. 1.2 Representação de um acontecimento ou conceito, sob uma forma susceptível de comunicação, de interpretação ou de tratamento, quer manualmente, quer por meios automáticos. 2. Grupo de caracteres alfabéticos, numéricos, alfanuméricos ou quaisquer outros, que representam uma condição ou valor específico. Os dados são, na realidade, os blocos construtivos da informação. 2.1 Uma referência não-elaborada, algo não-interpretado, não-classificado, não-estruturado, não-ajustado a um contexto [...]. 2.2 Informação em forma codificada. 3. "Fatos, noções ou instruções representados de uma forma conveniente para um processo de comunicação, uma interpretação ou um processamento quer humano, quer através de meios automáticos. Os dados analógicos são representados por funções contínuas, enquanto que os dados digitais são representados por funções discretas" (CUNHA; CAVALCANTI, 2008, p.113).

Davenport (1998, p. 19) define dados como "observações sobre o estado do mundo'". Dados brutos ou entidades quantificáveis podem ser trabalhados por pessoas ou pelas tecnologias da informação. Setzer (1999) destaca enquanto características de dado:

> [...] é uma seqüência de símbolos quantificados ou quantificáveis. [...] Como são símbolos quantificáveis, dados podem ser armazenados em um computador e processados por ele. [...] um dado é necessariamente uma entidade matemática e, desta forma, puramente *sintática* (SETZER, 1999, p. 1).

Os dados são definidos como fatos em uma forma primária e não necessariamente são compostos por fatos alfanuméricos (STAIR, 1998; O'BRIEN, 2003). Pode-se considerar um texto como um dado, já que os alfabetos são símbolos e podem ser quantificáveis como um conjunto finito constituindo uma base numérica (STAIR, 1998; SETZER, 1999; O'BRIEN, 2003). Também constituem dados os elementos do tipo áudios gravados, animações, figuras, fotos que podem ser quantificáveis e apresentar dificuldade ao distinguir a reprodução a partir da representação quantificada com os originais (STAIR, 1998).

O dado isolado de qualquer elemento em sua forma bruta apresenta significado restrito e pouco informativo, porém, quando atribuído em um contexto, um conjunto de dados pode gerar informações (FELIX, 2003; O'BRIEN, 2003). Os dados brutos perpassam pela eta-

pa de processamento, que está relacionada à intervenção humana. Durante o processamento, geram as informações de forma a ganhar um novo sentido em um determinado contexto, considerados, assim, produtos acabados (FELIX, 2003).

De maneira isolada, dados podem representar um estado, coisa ou evento através de conjuntos, fatos ou situações que apresentam valor definido ou significado (FELIX, 2003). O processamento de dados em um computador limita-se exclusivamente a manipulações estruturais realizadas por meio de *softwares* através de funções matemáticas (SETZER, 1999). Os dados, em sistemas de informação, são mais que matéria-prima. Eles são considerados um poderoso e valioso recurso organizacional utilizado pelos gestores e profissionais da área das tecnologias da informação (O'BRIEN, 2003).

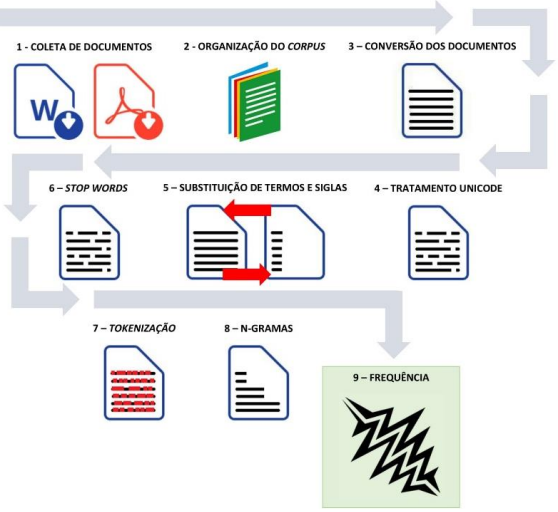
Setzer (1999) enfatiza a impossibilidade de transferir conhecimento entre indivíduos sem interação pessoal entre os envolvidos através de vivência e experiência. O que se transmite são dados que devem representar, da melhor maneira possível, as informações obtidas a partir deles.

3 METODOLOGIA

Esta pesquisa se classifica, quanto à finalidade/natureza, como aplicada; quanto à abordagem do problema, como quali-quantitativa; quanto aos objetivos, como exploratória (GIL, 2010).

A construção do referencial teórico foi realizada por meio da leitura de artigos científicos e livros disponibilizados pelo Portal de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES, Google Scholar e Google Books. Já as etapas referentes aos dados empíricos da pesquisa foram adaptadas de McKinney (2018) conforme apresentado na Figura 1:

Figura 1 – Fluxo para extração de frequência de termos



Fonte: Elaborado pelos autores.

Contemplam as etapas referentes ao percurso realizado para se alcançar os resultados dos dados empíricos:

a) Interação com o mundo externo (etapa 1): coleta de artigos científicos e resumos expandidos (restringindo-se a textos publicados em língua portuguesa) dos anais do Encontro Nacional de Pesquisa em Ciência da Informação – ENANCIB. Os da-

dos foram coletados entre os dias 27 de abril e 01 de maio de 2018 e referem-se ao período entre 2012 e 2017. No ano de 2019, foram acrescidos ao *corpus* os artigos completos e resumos expandidos publicados no ano de 2018;

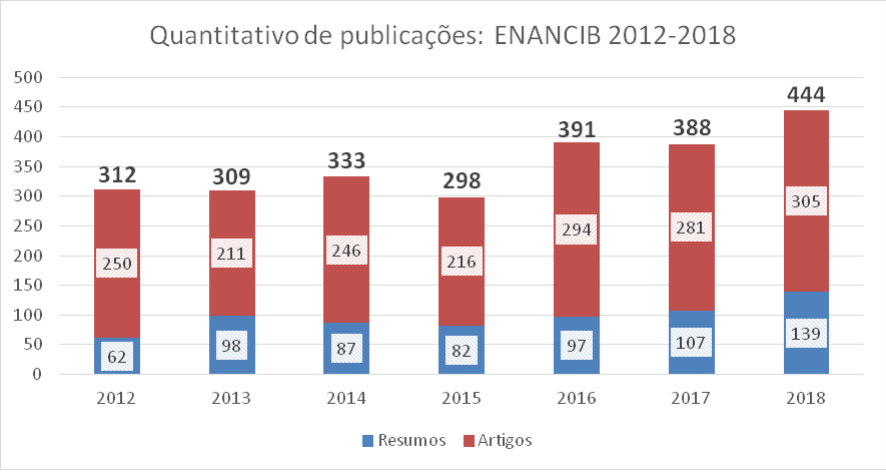
b) Preparação e pré-processamento (etapas 2, 3, 4, 5 e 6): organização, limpeza, manipulação, combinação, normalização, tratamento dos dados;
c) Transformação (etapas 7 e 8): operações matemáticas e estatísticas aplicadas em conjuntos de dados a fim de obter resultados significativos realizada nas etapas 7 e 8;
d) Apresentação (etapa 9): apresentação dos resultados.

Para alcançar os resultados, foram desenvolvidos algoritmos utilizados o *framework* Jupyter[1] junto com a linguagem de programação Python[2] e as bibliotecas PDFMiner[3], NLTK[4] e plotly[5]. Os algoritmos para extração de frequência dos termos bem como o gráfico dinâmico[6] contendo o comportamento de termos de composição estão disponibilizados através do GtiHub[7].

4 RESULTADOS E DISCUSSÕES

A composição do *corpus* foi realizada através de documentos do tipo artigos completos e resumos expandidos publicados entre os anos de 2012 e 2018 nos anais do Encontro Nacional de Pesquisa em Ciência da Informação - ENANCIB, principal evento brasileiro na área da Ciência da Informação. Foram coletados um total de 2.475 documentos, sendo 1803 artigos completos e 672 resumos expandidos distribuído entre o intervalo analisado conforme apresentado no Gráfico 1.

Gráfico 1 – Quantitativo anual de artigos completos e resumos expandidos publicados por anais do ENANCIB 2012-2018



Fonte: Elaborado pelos autores.

[1] Jupyter. Disponível em: https://jupyter.org/. Acesso em: 16/11/2020.
[2] Python. Disponível em: https://www.python.org/. Acesso em: 16/11/2020.
[3] PDFMiner. Disponível em: https://pypi.org/project/pdfminer/. Acesso em: 16/11/2020.
[4] NLTK - *Natural Language Toolkit*. Disponível em: https://www.nltk.org/. Acesso em: 16/11/2020.
[5] plotly. Disponível em: https://plotly.com/. Acesso em: 16/11/2020.
[6] Algoritmo do gráfico dinâmico do comportamento dos termos de composição. Disponível em: https://bit.ly/3iugwpi.
[7] Github. Plataforma de desenvolvimento colaborativo para hospedar, revisar códigos, gerenciar projetos e criar *software* de maneira colaborativa. Algoritmos e resultados disponíveis em: https://bit.ly/2YTrJIs.

Os documentos coletados, em formato PDF, que somados alcançaram mais de 2GB de dados, foram convertidos[8] para o formato TXT[9], legível pelo algoritmo, que resultou em uma redução no tamanho dos arquivos para 94MB. Após este processo, foi realizado o descarte de documentos que apresentaram problemas de Unicode[10] devido a erros gerados no processo de conversão de documentos PDF contendo imagens ou com restrições de privacidade para texto, sendo: 6 documentos referentes ao ano 2013; 18 documentos referentes ao ano 2015; 2 documentos referentes ao ano 2016 e; 1 documento referente ao ano de 2017.

A conversão de termos em siglas ou vice-versa foi realizada por meio de Expressões Regulares de maneira que as frequências dos termos com o mesmo significado não fossem divididas, como, por exemplo, a sigla BD foi convertida para Bando de Dados. Posteriormente, todas as *stop words* – palavras de parada - foram excluídas do *corpus* por meio da biblioteca NLTK. Além da lista padrão disponibilizada pela biblioteca, foi adicionado, junto ao algoritmo, outros termos com as mesmas características das *stop words*. São exemplos de palavras de parada: “e”, “de”, “assim” e “como”.

A *tokenização* que divide os textos contidos no *corpus* em frases, palavras, símbolos e outros elementos convertidos em uma sequência de palavras separadas por tratamento de pontuação e espaçamento foi realizada por meio da biblioteca NLTK. Ainda, na fase de transformação dos dados, foi executada a função *n*-grama que criou um total de 6.640.564 unigramas, 6.638.116 bigramas e 6.635.668 trigramas de todo o *corpus*, sendo exportadas, ao final, listas de frequências contendo os mil primeiros termos de cada tipo de *n*-grama, considerando, assim, a qualidade dos termos. Exemplos de bigrama e trigrama com características de baixa qualidade são “dado_email” e “nome_dado_trabalho”.

As frequências dos termos dos tipos de *n*-grama extraídas de *corpora* de documentos permitem identificar o comportamento diacrônico dos termos quando comparados em um determinado intervalo de tempo. A classificação de comportamento de termos contribui para um melhor entendimento e uso de termos extraídos da produção científica apresentando, assim, caminhos norteadores para a área estudada (SOUZA, 2020).

Considera-se, nesta pesquisa, que os termos extraídos de *corpus* apresentam os seguintes comportamentos e características:

a) comportamento contínuo – assiduidade, que não apresenta interrupções de frequência ao longo do período analisado, podendo ser: a) regular – apresenta regularidade de frequência do termo com variações menores que 100%; b) irregular – apresenta irregularidade de frequência com variações maiores que 100% para cima e para baixo; c) ascensão – apresenta aumento de frequência do termo ao longo do período analisado, superior a 100%; e d) descensão – apresenta queda de frequência do termo ao longo do percurso, superior a 100%;
b) comportamento inconstante – apresenta ausência de frequência de termos no intervalo analisado: a) regular – apresenta regularidade de frequência em um determinado período analisado e que não ultrapassa 100%; b) irregular – apresenta irregularidades entre os anos analisados, onde a frequência ultrapassa 100% para mais e para menos; c) ascensão – apresenta uma crescente na frequência do termo nos últimos

[8] Algoritmo de conversão de documentos. Disponível em: https://bit.ly/2D19vfr.
[9] TXT - Ficheiros de texto que podem conter texto simples, sem formatação.
[10] Unicode - Padrão que permite aos computadores representar e manipular, de forma consistente, texto de qualquer sistema de escrita existente.

anos analisados ou período; e d) descensão – apresenta queda de frequência dos últimos termos analisados ou período.

O termo dado apresentou frequência de 27.879, extraídos de 2.448 documentos científicos. No ano de 2012, o termo dado apareceu 2.824 vezes, enquanto no ano de 2013, o termo apresentou frequência de 2.896, o que representa um aumento de 3% em relação ao ano anterior. No ano de 2014, o termo apresentou frequência de 3.682, o que equivale a um aumento de 27%. Já no ano de 2015, houve uma queda no quantitativo de termos em -18%, resultando em uma frequência de 3.019. Os anos seguintes apresentaram aumentos de frequência, sendo 4.028 ou 33% no ano de 2016, 5.606 ou 39% no ano de 2017 e 5.824 ou 4% no ano de 2018, conforme apresentado no Gráfico 2

Gráfico 2 – Comportamento do termo dado

Fonte: Elaborado pelos autores.

Entre os anos de 2012 e 2018 o termo dado alcançou uma diferença acumulada de frequência equivalente a 106%, o que permite refletir que houve um aumento de mais que o dobro do uso do termo nas produções científicas da área da Ciência da Informação. Entretanto, o termo dado no *corpus* apresenta diferentes composições de termos, dos tipos bigramas e trigramas. A Figura 2 apresenta 26 composições dos termos com maiores frequências extraídas dos artigos completos e resumos expandidos gerados a partir do termo raiz.

Figura 2 – Termos de composição a partir do termo dado

Fonte: Elaborado pelos autores.

Os termos compostos, no acumulado de frequência, apresentam menores valores quando comparado à frequência do termo raiz, como, por exemplo, os bigramas "base_dados", que apresenta uma frequência de 3.345; "dados_pesquisa", com uma frequência de 2.062; e "coleta_dados" com uma frequência de 1.744. Contudo, a existência de um menor volume de termos compostos, a partir do termo raiz, apresentam valores significativos e que permitem realizar o mapeamento científico do termo. Numa lista contendo 25 termos formados por composição a partir do termo dado, entre bigramas e trigramas, alcança-se um total de frequência de 14.739, representando 52,86% da frequência do termo conforme apresentado no Quadro 1.

Quadro 1 – Termos composição e frequências.

base_dados bases_dados,3345; dados_pesquisa,2062; coleta_dados1744; fonte_dados,1409; fonte_dados_pesquisa,938; banco_dados,890; dados_coletados,718; dados_abertos,696; dados_informação, 645; conjunto_dados,625; dados_obtidos,294; instrumento_coleta_dados,258; dados_científicos,180; ciclo_vida_dados,154; dados_web,132; dados_bibliográficos,128; gestão_dados,123; dados_governamentais,107; mineração_dados,67; dados_quantitativos,57; dados_informação_conhecimento,47; dados_altmétricos,40; dados_interligados,29; modelagem_dados,26; dados_imagens,25.

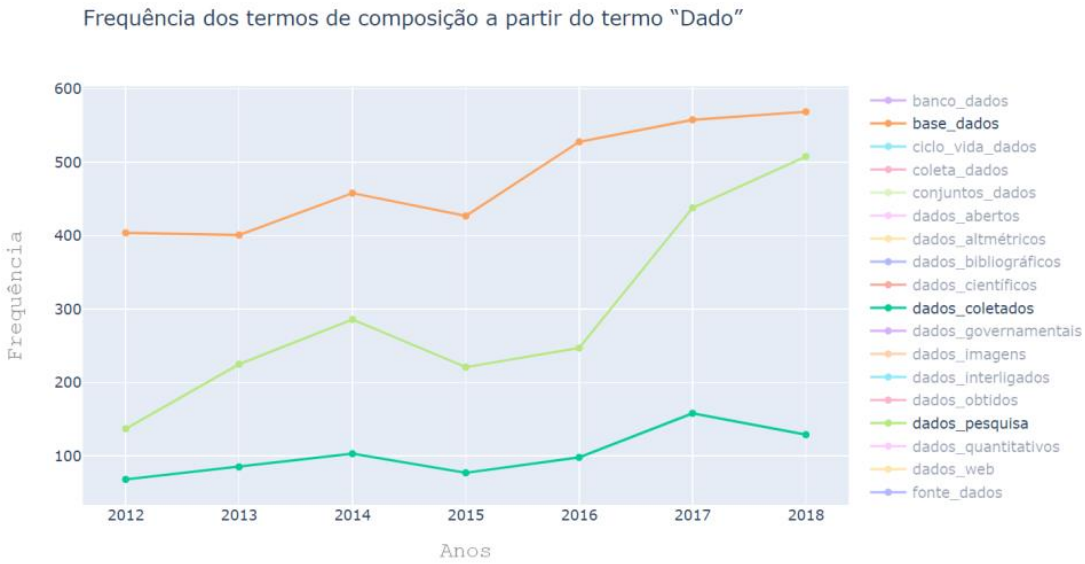
Fonte: Elaborado pelos autores.

Dentre os termos compostos destaca-se "base_dados", representado pela cor laranja, que apresentou frequência de 404 no ano de 2012; 401 em 2013, representando queda de -1%; 458 em 2014, significado aumento de 14%; 427 em 2015, representando queda de -7%; 528 em 2016, ou crescimento de 24%; 558 em 2017, ou crescimento de 6% e; 569 em 2018, representando crescimento de 2%. Toda alteração de comportamento do termo no que se refere ao crescimento ou queda está relacionado ao ano anterior. Numa visão geral, o termo "base_dados" apresentou crescimento de 41% entre os anos de 2012 a 2018.

O termo "dados_pesquisa", representado pela cor verde claro, apresentou frequência de 137 no ano de 2012; 225 em 2013, apresentando crescimento de 64%; 286 em 2014, que significa crescimento de 27%; 221 no ano de 2016, o que significa queda de -12%; 438 no ano de 2017, apresentando, assim, um crescimento de 77% e; 508 no ano de 2018, representando um crescimento de 16%. Assim como no termo anterior, toda alteração de comportamento do termo está relacionada ao ano anterior analisado. O termo "dados_pesquisa" apresentou crescimento de 271% no período analisado.

Já o termo "dados_coletados", representado pela cor verde escuro, apresentou frequência de 68 no ano de 2012; 85 em 2013, apresentando crescimento de 25%; 103 em 2014, que significa um acréscimo de 25%; 77 em 2015, apresentando queda de -25%; 98 em 2013, novamente apresentando crescimento de 27%; 158 em 2017. Nesse ano, o termo apresentou o maior crescimento no período analisado, representando aumento de 61% em relação a 2012. No ano de 2018, a frequência do termo teve um decréscimo de 18%, alcançando 129 menções. Toda alteração de comportamento do termo está relacionada ao ano anterior. O termo "dados_coletados" apresentou crescimento de 90% entre os anos de 2012 a 2018. O comportamento dos termos é apresentado no Gráfico 3.

Gráfico 3 – Comportamento de termos de composição



Fonte: Elaborado pelos autores.

Os termos "base_dados", "dados_pesquisa" e "dados_coletados" apresentam comportamentos similares ao longo dos anos. Um dos pontos de destaque é o ano de 2015, onde os termos apresentaram queda em suas respectivas frequências, conforme Gráfico 3. Infere-se que este decréscimo possa ter acontecido devido ao menor número de publicações científicas no ano de 2015, embora o quantitativo de pesquisas não possua relação direta como o número de frequências dos termos e sim, com o conteúdo das pesquisas científicas. De maneira oposta, percebe-se que, no ano de 2018, a frequência do termo "dados_coletados" apresenta queda de -18% em relação ao ano anterior, apesar de possuir o maior número de publicações.

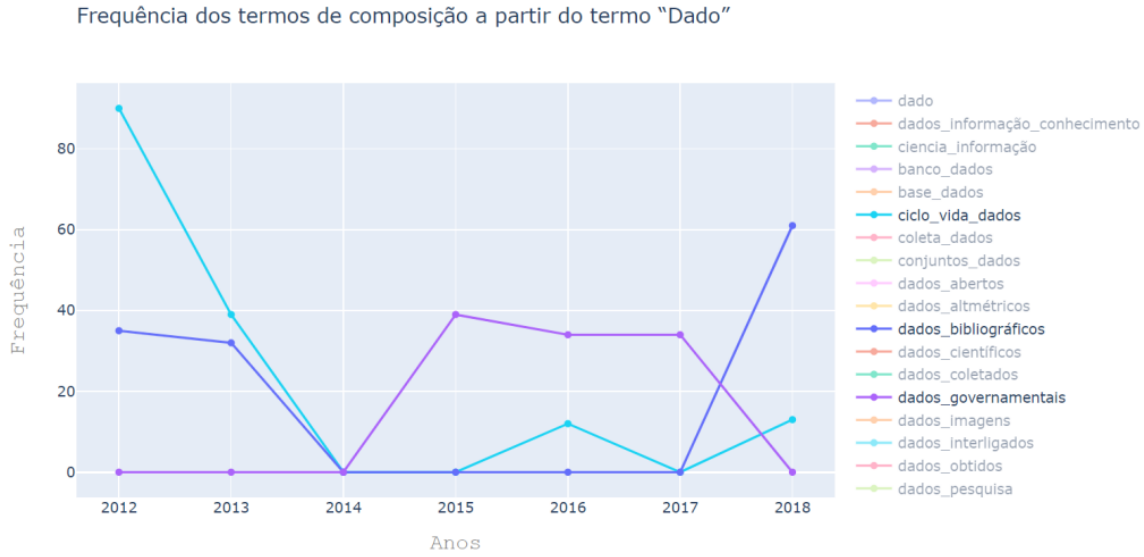
Destaca-se que, dentre os termos analisados, considerando que se possa ter mais de um comportamento ao longo do período analisado, o termo "base_dados" possui comportamento contínuo, com frequência regular em todos os períodos analisados, com variações que não ultrapassam 100% para mais ou para menos. O termo "dados_pesquisa" também apresenta um comportamento contínuo, entretanto, destaca-se como característica, a ascensão do termo com percentuais acumulados que alcançam 271%, entre o primeiro e o último ano analisado, além de apresentar uma linha crescente ao longo do período. O termo "dados_coletados" apresenta as mesmas características do termo "base_dados", portanto, uma característica regular.

O termo "ciclo_vida_dados", representado pela cor azul claro, apresentou frequência de 90 no ano de 2012; 39 em 2013, apresentando queda de -57%. Nos anos 2014 e 2015, o termo apresentou frequência igual a zero ou frequência após o milésimo termo, que não são consideradas neste estudo devido a qualidade dos termos extraídos do *corpus*. O ano de 2016 apresentou frequência 12 e, novamente, em 2017, o termo apresentou frequência igual a zero. Em 2018, o termo teve frequência igual a 13.

O termo "dados_bibliográficos", representado pela cor azul, apresentou frequência de 35 no ano de 2012. Em 2013, o termo apresentou queda de 9%, totalizando 32 menções. Nos anos de 2014 a 2017, o termo apresentou frequência igual a zero ou após o milésimo termo. Em 2018, o termo teve frequência de 61.

Já o termo “dados_governamentais”, representado pela cor lilás, apresentou as seguintes frequências entre os anos de 2015 a 2017, respectivamente: 39, 34 e 34, equivalente a uma queda de -13%. Nos demais anos, o termo não apresentou frequência ou apresentou frequência após o milésimo termo extraído do *corpus,* conforme apresentado no Gráfico 4.

Gráfico 4 – Comportamento de termos de composição



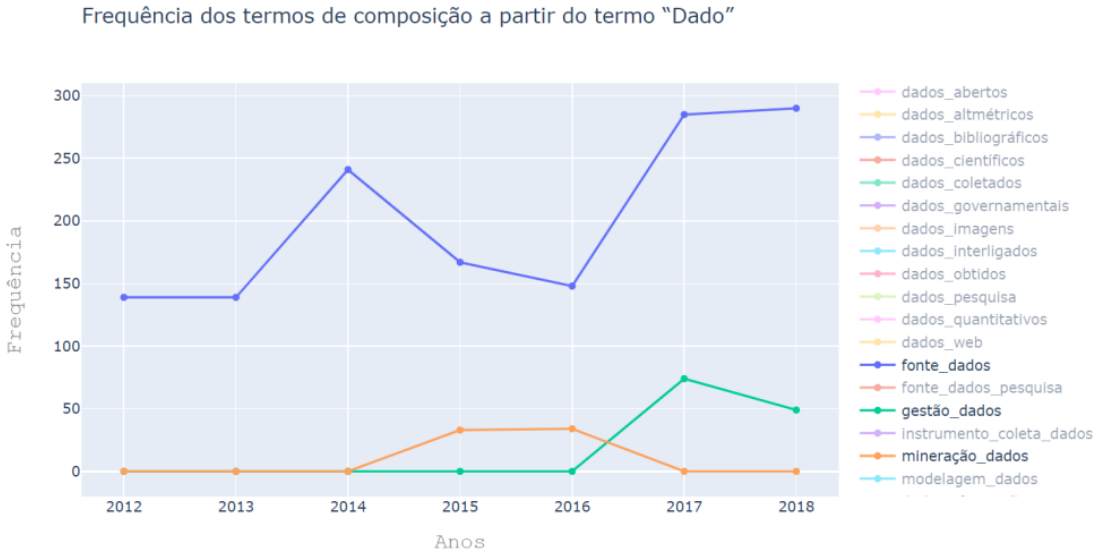
Fonte: Elaborado pelos autores.

Assim, os três termos apresentados, “ciclo_vida_dados”, “dados_bibliográficos” e “dados_governamentais” apresentaram comportamentos inconstantes, com características regulares como a ausência de frequências entre os intervalos analisados. Pode-se considerar que os termos tenham surgido esporadicamente para suprir lacunas da ciência ou que estejam relacionados a grupos de pesquisadores ou pesquisas específicas.

O termo “fonte_dados”, representado pela cor azul, apresentou frequência de 139 nos anos de 2012 e 2013. Em 2014 houve crescimento de 73%, com frequência de 241. No ano de 2015, o termo apresentou queda de -31%. Em 2016, foram 148 menções, significando queda de -11%. Já nos anos de 2017 e 2018, o termo apresentou crescimento sendo, respectivamente, 285, com aumento de 93% e 290, com aumento de 2%. A alteração de comportamento, seja para crescimento ou queda, está relacionada ao ano anterior analisado.

O termo “gestão_dados”, representado pela cor verde, não apresentou frequência entre os anos de 2012 a 2016. Em 2017, o termo apresentou frequência de 74 e no ano de 2018, 49, equivalente a uma queda de -34% em relação ao ano anterior. Já o termo “mineração_dados”, representado pela cor laranja, apresentou frequência de 33 no ano de 2015 e 34 no ano de 2016, resultando em um crescimento de 3%, conforme apresentado no Gráfico 5. Neste trabalho, os termos que apresentaram frequência igual a zero não apareceram entre os resultados extraídos do *corpus* ou podem aparecer após o milésimo termo extraído de cada categoria de *n*-grama.

Gráfico 5 – Comportamento de termos de composição



Fonte: Elaborado pelos autores.

O termo "fonte_dados" apresentou comportamento contínuo em todos os anos analisados, além da característica regular, não ultrapassando, assim, mais de 100% para mais ou para menos entre os anos analisados. Apresentou, ainda, ascensão de 109% no acumulado de frequência entre os anos de 2012 a 2018.

O termo "gestão_dados" apresenta comportamento irregular, com ausência de menções entre os anos de 2012 e 2016, e uma regularidade nos dois últimos anos analisados, o que se supõe que o termo esteja ganhando notoriedade na comunidade científica.

Já o termo "mineração_dados" que possui comportamento irregular, apareceu em pesquisas da área da Ciência da Informação nos anos de 2015 e 2016. Nos anos seguintes, o termo apresentou descensão.

Considera-se que, um mesmo termo, pode apresentar uma combinação de características quando analisados períodos dentro de um intervalo. Por exemplo, o termo "ciclo_vida_dados" que apresenta comportamento irregular com descensão entre 2012 e 2014 e ascensão entre 2017 e 2018.

Um conjunto de fatores pode ser determinante para a mudança de comportamento dos termos ao longo de um determinado período analisado, sendo alguns dos exemplos:

a) aumento ou redução no quantitativo de pesquisadores e, consequentemente, no quantitativo de pesquisas publicadas junto ao ENANCIB;
b) nova temática apresentada a cada edição do evento;
c) surgimento de novos programas de pós-graduação, podendo ocasionar aumento do quantitativo de publicações junto ao evento;
d) diferentes políticas dos Grupos de Trabalho – GTs do evento, onde não existe um quantitativo mínimo e máximo de pesquisas aprovadas por edição;
e) fator interdisciplinaridade da Ciência da Informação, recebendo pesquisas de áreas diferentes;
f) dificuldade em quantificar termos por GT, uma vez que um determinado termo pode ser mencionado em um ou mais GTs.

5 CONSIDERAÇÕES FINAIS

O termo dado, que constitui um dos elementos da tríade dado, informação e conhecimento, estudado na Ciência da Informação, apresentou comportamento contínuo e ascendente ao longo do período analisado, apresentado um acúmulo de frequências igual a 27.879 e média anual de 3.983. Esses números foram extraídos de 2.448 documentos do tipo artigo completo e resumo expandido, representando crescimento de 106% no intervalo analisado.

Embora seja um termo em voga na área da Ciência da Informação, o termo dado possui características generalistas, bem como pouca representatividade quando analisado na sua forma bruta, conforme embasamento teórico, entretanto, o real comportamento do termo dado tem se apresentado através de termos formados por composição e que abordam assuntos com maiores especificidades relacionados à área do *corpus* estudado, tais como "base_dados", com frequência total de 3.345, "dados_pesquisa" com 2.062 menções, "coleta_dados" com 1.774 de frequencia, "fonte_dados" com 1.409 e "banco_dados" 890 extrações. As somas das frequências dos termos formados por composição alcançam 52,86% do total de frequência do termo raiz que é de 27.879.

Os termos de composição do tipo bigramas e trigramas podem apresentar diferentes comportamentos que são reflexos de pesquisas científicas realizadas na área. Dentre os comportamentos constam o contínuo e inconstante e as características regular, irregular, ascensão e descensão. Considera-se que, dentre os comportamentos, os do tipo contínuo e regular, são termos estabilizados na linguagem de domínio, como apresentado pelo comportamento do termo "base_dados".

Os termos irregulares em ascensão caracterizam-se pela possibilidade do surgimento de novos termos no domínio da linguagem a serem explorados, como apresentado pelo comportamento do termo "gestão_dados", que apresenta frequência somente nos anos de 2017 e 2018. Comportamento contrário é encontrado no termo "mineração_dados" que apresenta irregularidade e descensão a partir de 2017, não apresentando frequência ou apenas frequência após o milésimo termo extraído.

Enquanto contribuição para a Ciência da Informação, a pesquisa realizou o mapeamento científico dos termos de composição formados a partir do termo raiz, apresentando, assim, quais são os termos em uso no domínio da linguagem, bem como os termos mais utilizados.

O pressuposto da pesquisa foi confirmado ao constatar que o comportamento do termo dado não remete unicamente a realidade de sua frequência, uma vez que termos de composição, que compõe o mapeamento científico do termo, são incluídos no percentual de frequência do termo raiz. Os termos formados por composição apresentam outros significados de relevância para o domínio da linguagem explorado, porém, diferentes do termo original.

Sugere-se, para pesquisas futuras, o estudo dos comportamentos dos termos informação e conhecimento, também em *corpus* da Ciência da Informação, bem como um estudo sobre o comportamento dos termos da tríade conjuntamente e dos termos de transição entre as áreas como "dado_informação", "informação_conhecimento" e "conhecimento_dado".

REFERÊNCIAS

BOISOT, M. **Competitive advantage in the information economy**. Oxford; New York: Oxford University Press, 1988.

BORKO, H. Information science: what is it? **American Documentation**, v.19, n.1, p.3-5, 1968.

CUNHA, M. B.; CAVALCANTI, C. R. O. **Dicionário de Biblioteconomia e Arquivologia**. Brasília: Briquet de Lemos, 2008.

DATA. *In*: OXFORD Advanced Learner's Dictionary of Current English. Oxford: Oxford University Press, 2005.

DATA. *In*: ONLINE Dictionary for Library and Information Science. Santa Barbara: ABC-CLIO, 2020. Disponível em: https://products.abc-clio.com/ODLIS/odlis_d.aspx/. Acesso em: 01 nov. 2021.

DAVENPORT, T. E. **Ecologia da informação:** por que só a tecnologia não basta para o sucesso na era da informação. São Paulo: Futura, 1998.

GIL, A. C. **Como elaborar projetos de pesquisa**. 5. ed. São Paulo: Atlas, 2010.

FELIX, W. **Introdução à Gestão da Informação.** Campinas: Alínea, 2003.

MCKINNEY, W. **Python para análise de dados**: tratamento de dados com pandas, numpy e ipython. São Paulo: Novatec, 2018.

O'BRIEN, J. A. **Sistemas de Informação e as decisões gerenciais na era da internet**. 9. ed. São Paulo: Saraiva, 2003.

PASCHOALIN, M. A.; SPADOTO, N. T. **Gramática**: teoria e exercícios. São Paulo: FTD, 1996.

PINHEIRO, L.V.R. Processo evolutivo e tendências contemporâneas da Ciência da Informação. **Informação & Sociedade**: Estudos. João Pessoa, v. 15, n.1, p.13-48, 2005.

RIO-TORTO, G. M. Mecanismos de produção lexical no português europeu. **Alfa**, v.42, n.esp, p. 15-32, 1998. Disponível em: https://periodicos.fclar.unesp.br/alfa/issue/view/298. Acesso em: 01 nov. 2020.

SARACEVIC, T. Interdisciplinarity nature of Information Science. **Ciência da Informação**, Brasília, v. 24, n. 1, p. 36-41, 1995.

SARACEVIC, T. Ciência da Informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 1, n. 1, p. 41-62, 1996.

SEMIDÃO, R. A. M. Dados, Informação e Conhecimento: elementos de análise conceitual. **DataGramaZero - Revista de Informação**, Rio de Janeiro, v. 14, n. 4, p. 10, 2013. Disponível em: https://brapci.inf.br/index.php/article/download/52967. Acesso em: 01 nov. 2020.

SEMIDÃO, R. A. M. **Dados, informação e conhecimento enquanto elementos de compreensão do universo conceitual da ciência da informação**: contribuições teóricas. 2014. 198 f. Dissertação (Mestrado em Ciência da Informação) - Universidade Estadual Paulista, Faculdade de Filosofia e Ciências de Marília, 2014. Disponível em: http://hdl.handle.net/11449/110783/. Acesso em: 01 nov. 2020.

SETZER, V. Dado, informação, conhecimento e competência. **DataGramaZero – Revista de Ciência da Informação**, Rio de Janeiro, n. 0, dez. p. 1-14, 1999. Disponível em: https://www.ime.usp.br/~vwsetzer/datagrama.html. Acesso em: 01 jul. 2020.

SIRIHAL, A. B.; LOURENÇO, C. A. Informação e conhecimento: aspectos filosóficos e informacionais. **Informação & Sociedade**, João Pessoa, v. 1, n. 12, p. 1–15, 2002.

SOUZA, M. **O comportamento de termos da Ciência da Informação por meio da modelagem de tópicos**. 2020. 404 f. Tese (Doutorado em Gestão e Organização do Conhecimento) – Programa de Pós-Graduação em Gestão e Organização do Conhecimento, Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2020. Disponível em: https://repositorio.ufmg.br/handle/1843/34292. Acesso em: 01 nov. 2020.

STAIR, Ralph M. **Princípios de Sistemas de Informação:** uma abordagem gerencial. 2. ed. Rio de Janeiro: Ltc, 1998.

SUKKARIEH, J. Z.; PULMAN, S. G.; RAIKES, N. Auto-marking: using computational linguistics to score short, free text responses. *In*: ANNUAL CONFERENCE OF THE INTERNATIONAL ASSOCIATION FOR EDUCATIONAL ASSESSMENT, 29, 2003, Manchester. **Proceedings**... [S.l.]: IAEA, 2003. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.7417&rep=rep1&type=pdf. Acesso em: 01 nov. 2020.