5º WIDaT

ANAIS do Workshop de Informação, Dados e Tecnologia

05 e 06 de dezembro/2022

UFES - Campus Goiabeiras
Auditório da Biblioteca Central

Realização:

Universidade Federal do Espírito Santo

Centro de Ciências Jurídicas e Econômicas
Universidade Federal do Espírito Santo

PPGCI Programa de Pós-Graduação em Ciência da Informação

Apoio:

UnB

SIB Sistema Integrado de Bibliotecas

Tainacan

Departamento de Biblioteconomia
UFES

SEAD UFES

FAPES
Fundação de Amparo à Pesquisa e Inovação do Espírito Santo

Governo do Estado do Espírito Santo
Secretaria da Ciência, Tecnologia, Inovação, Educação Profissional e Desenvolvimento Econômico

ISBN: 978-65-00-61990-4

WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA
Universidade Federal do Espírito Santo – UFES
05 e 06 de dezembro de 2022, Vitória – ES

ANAIS WIDAT 2022

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO
UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

Organizadores
Henrique Monteiro Cristovão (PPGCI/UFES)
Daniela Lucas da Silva Lemos (PPGCI/UFES)

Vitória-ES
2023

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Central da Universidade Federal do Espírito Santo, ES, Brasil)

W924

Workshop de informação, dados e tecnologia (5. : 2022 dez. : Vitória).
Workshop de informação, dados e tecnologia (WIDAT 2022): anais do V WIDaT [recurso eletrônico] / Henrique Monteiro Cristovão, Daniela Lucas da Silva Lemos, organizadores. - Dados eletrônicos. - Vitória : PPGCI/UFES, 2022.
247 p. ; il.

Workshop realizado no período de 5 e 6 de dez. de 2022.
Inclui bibliografia.
ISBN: 978-65-00-61990-4
Modo de acesso: https://widat2022.ufes.br/anais

1. Ciência da informação - Workshop. 2. Tecnologia. I.Henrique Monteiro Cristovão. II. Daniela Lucas da Silva Lemos. III. Título.

CDU: 02

Elaborado por Kênia Aparecida Limas dos Santos – CRB-6 MG003236/O

Organização WIDaT 2022

Coordenação Geral:
Daniela Lucas da Silva Lemos (PPGCI/UFES)

Comissão Organizadora:
Dalton Lopes Martins (PPGCInf/UNB)
Gleice Pereira (PPGCI/UFES)
Henrique Monteiro Cristovão (PPGCI/UFES)
Marta Leandro da Mata (PPGCI/UFES)
Morgana Andrade (Biblioteca Central/UFES)
Paula Regina Ventura Amorim Gonçalez (Depto. Biblioteconomia/UFES)

Coordenação da Comissão Científica:
Henrique Monteiro Cristovão (PPGCI/UFES)

Comissão Científica:
Alisson Marques da Silva (CEFET-MG)
Ana Carolina Simionato (PPGCI/UFSCar)
Dalton Martins (PPGCInf/UNB)
Daniela Lucas da Silva Lemos (PPGCI/UFES)
Diana Vilas Boas Souto Aleixo (DARQ/UFES)
Douglas Dyllon Jerônimo de Macedo (CIN/UFSC)
Eduardo Ribeiro Felipe (UNIFEI)
Fabrício Mendonça (DCC/ UFJF)
Fernanda Farinelli (PPGCInf/UNB)
Jeanne Louize Emygdio (PUC-MG)
Jose Eduardo Santarem Segundo (PPGCI/UNESP/ USP)
Luciana Itida Ferrari (DARQ/UFES)
Marta Leandro da Mata (PPGCI/UFES)
Michel Pires (CEFET-MG)
Moisés Lima Dutra (CIN/UFSC)
Morgana Andrade (Biblioteca Central /UFES)
Patrícia Dias (UEMG)
Paula Regina Ventura Amorim Gonçalez (DBIB/UFES)
Renato Rocha Souza (CPDOC/FGV)
Sandro Rautenberg (PGCIN-UFSC)
Tânia Barbosa Salles Gava (DARQ/UFES)
Thiago Magela Rodrigues Dias (CEFET-MG)
Tiago Alves de Oliveira (CEFET-MG)
Washington Segundo (IBICT)

Comissão Técnica:
Bruna Stefane de Freitas (UFES)
Calíope Victor Spíndola de Miranda Dias (UnB)
Dirceu Flávio Macedo (Grupo TOIC/UFES)
Kris Ellen das Neves Teixeira (UFES)
Luísa Vernersbach Varejão (UFES)
Rafael Denerson Ramos de Sousa (UFES)
Silvana Pires Rocha Nogueira (UFES)

Sumário

A APLICABILIDADE DA FOLKSONOMIA NOS ESTUDOS ALTMÉTRICOS

THE APPLICABILITY OF FOLKSONOMY IN ALTMETRIC STUDIES

Skrol Salustiano[1], Fabio Castro Gouveia[(2)]
(1) IBICT-UFRJ, Rio de Janeiro, skrol@ufrj.br
(2) Fiocruz, Rio de Janeiro, fabio.gouveia@fiocruz.br

Resumo
A popularização e crescimento das redes sociais abriu novas possibilidades para a pesquisa científica. Nos estudos métricos da informação possibilitou o surgimento de novos campos de estudos como a Folksonomia, utilizada para etiquetagem social e a Altmetria, cuja premissa é medir a disseminação de documentos científicos nas redes sociais. **Objetivo:** Identificar a convergência entre os saberes e, principalmente, se é possível construir um modelo para o desenvolvimento de indicadores altmétricos. **Metodologia:** Levantamento bibliográfico e de dados realizada no Twitter, onde foram coletadas publicações que tivessem a hashtag "*openaccess*". O conteúdo foi analisado para identificar, principalmente, se existiam links para produções científicas. Dessa forma, tornou-se possível compreender a dinâmica da utilização das hashtags no ambiente acadêmico e como esses resultados podem ser aplicados nos Estudos Métricos da Informação. **Resultado:** Com os tweets capturados, analisados e classificados foi possível extrair informações sobre o tipo de conteúdo compartilhado e as áreas do saber predominantes. **Considerações Finais:** O dataset permitiu observar a possibilidade de construção de um modelo para identificar a convergência entre os saberes e, principalmente, se é possível o desenvolvimento de indicadores altmétricos.

Palavras-chave: Folksonomia, Altmetria, Redes Sociais, Indexação, Estudos Métricos.

Abstract
The popularization and growth of social networks has opened new possibilities for scientific research. In information metrics studies it has enabled the emergence of new fields of study such as Folksonomy, used for social labeling, and Altmetrics, whose premise is to measure the dissemination of scientific documents on social networks. **Objective:** To identify the convergence between the two fields of knowledge and, especially, if it is possible to construct a model for the development of altmetric indicators. **Methodology:** Bibliographic and data survey conducted on Twitter, where publications that had the hashtag "openaccess" were collected. The content was analyzed to identify, mainly, if there were links to scientific productions. Thus, it became possible to understand the dynamics of the use of hashtags in the academic environment and how these results can be applied in Information Metrics Studies. **Result:** With the captured, analyzed and classified tweets it was possible to extract information about the type of shared content and the predominant knowledge areas. **Final Considerations:** The dataset allowed us to observe the possibility of building a model to identify the convergence between the knowledge areas and, especially, if it is possible to develop altmetric indicators.

Keywords: Folksonomy, Altmetrics, Social Networks, Indexing, Metric Studies.

1. Introdução

Nas últimas duas décadas, com a consolidação dos ambientes digitais surgiram novas oportunidades para o acompanhamento do fazer ciência, por meio de dados que sinalizam uma possível relevância e/ou interesse ao assunto abordado. Como na previsão de Bossy (1995), a internet se tornou um campo para observação da "ciência em ação".

Neste novo ambiente que se abriu para a pesquisa científica surgiram novas maneiras para organizar, representar e recuperar informações, baseadas em rótulos (tags), que servem para marcar e agrupar conteúdos similares. Guedes e Dias (2010) ao analisarem a nomenclatura apropriada para a indexação de objetos digitais, observaram que "um mesmo documento é acessado, manipulado, organizado e recuperado por uma infinidade de usuários

em diversas partes do mundo e ao mesmo tempo." (GUEDES; DIAS, 2010, 47), o que favorece a descentralização na organização das informações.

Assim, esse sistema de atribuição de tags ou etiquetagem digital dos conteúdos surge como uma adaptação aos modelos de taxonomia e passou a ser chamado de Folksonomia[1]. O diferencial é que dentro dessa nova dinâmica, as palavras-chave que serão indexadas passam a ser construídas e geridas pelos termos livres atribuídos pelos usuários ao compartilharem os conteúdos.

2. Objetivos

Com base nessa premissa, esse estudo ainda em andamento, apresenta os primeiros resultados, da pesquisa que busca analisar as potencialidades da Folksonomia aplicada aos Estudos Métricos da Informação (EMI), com foco na Altmetria, para verificar se teoricamente existe a convergência entre os saberes. Enquanto é feita essa avaliação, observar a possibilidade de construção de um modelo prático e funcional para os EMI. Para atingir os objetivos, a pesquisa teve sua metodologia baseada na revisão narrativa, que identificou e discutiu as inter-relações entre Altmetria e Folksonomia. Para isso foram utilizados como bibliografia estudos recorrentemente citados, incluindo autores nacionais e internacionais.

3. Procedimentos Metodológicos

A pesquisa teve dois momentos distintos: o primeiro foi a análise do conceito da Folksonomia, quando buscou-se identificar as possibilidades de aplicação para a organização do conhecimento. Nesta etapa foi realizada a revisão narrativa, para mapear os avanços realizados em mais de uma década de estudos da Folksonomia.

Na segunda parte, em consonância com o objetivo da pesquisa de identificar o impacto da participação pública na organização do conhecimento, foi definido como termo de levantamento a hashtag "#openacess".

O Twitter[2] foi escolhido como o ambiente mais adequado por permitir a coleta de conteúdos, com variedade de público e baixa restrições. A captura das informações foi realizada com a ferramenta TAGSExplorer[3], com período compreendido entre 20 de julho de 2022 a 2 de setembro de 2022. A janela temporal tinha como objetivo inicial completar o período de três meses. No entanto, ao atingir o volume de 200 mil tweets a ferramenta começou a apresentar erros na captura dos conteúdos. Dessa forma, para não prejudicar a qualidade do dataset a coleta foi interrompida antecipadamente, mas sem prejuízos à qualidade das informações coletadas que foram além do esperado.

Dentro da janela temporal foram extraídos 298.516 tweets ou publicações, que foram tratados e modelados na ferramenta Rstudio.

A primeira triagem ou visualização foi identificar quais eram as principais hashtags associadas ao conteúdo coletado. Foi identificado muitas publicações com apenas uma única hashtag e em sua maioria não tinham links associados. Estes foram excluídos, pois não entregaram o principal resultado da busca: qual tipo de informação estava sendo compartilhada nos links.

Em seguida, foram também eliminados da análise de usuários com menos de cinco mil seguidores. Essa opção aconteceu pela identificação deste público ter média de apenas quatro retuites. Como o TAGSExplorer não oferece outra métrica de engajamento (curtidas, salvamentos e cliques nos links), essa métrica serviu para apontar a dispersão, apropriação e envolvimento com os conteúdos publicados.

Após a limpeza dos dados, o dataset resultou em um conjunto 19.437 tweets, cujas hashtags das publicações totalizaram 174.933, o que representa uma média de nove hashtag por publicação.

Esse volume de tweets selecionados ainda passou por outros tratamentos realizados no próprio Rstudio, com o objetivo de extrair o tipo de publicação, posteriormente classificadas em: Links para

[1] Termo cunhado por Thomas Vander Wal (2007) que significa a junção das palavras: "folk" pessoas; e "taxonomia" = classificação. Ver também: Furner (2009)

[2] https://tags.hawksey.info

[3] https://twitter.com

Artigos, Links Diversos, Links Mistos e Sem Links.

Dentro dos processos descritos acima o principal foi o de descompactação das URLs encurtadas. Para o processo foram utilizados os pacotes DyplR e Longurl. O primeiro foi extrair as URLS das menções e copiar para uma nova coluna, como forma de facilitar a descompactação. Em seguida foi utilizado o pacote Longurl que extraiu as urls originais.

Com base nos links foi possível identificar as revistas científicas ou periódicos e novamente com o DyplyR criar a relação entre tipo de periódico e área do saber. Os endereços que não pertenciam a publicações científicas foram classificados como "Diversos", pois a maioria era informativos de eventos (Congressos, seminários, webinars, entre outros). Essa classificação dos links diversos foi realizada de forma manual: foram agrupados os links similares e posteriormente foram analisados.

4. Resultados

Com base nos dados coletados observou-se que a maior parte do conteúdo foi para o compartilhamento de links. Deste volume, conforme pode ser observado no Apêndice A, do Gráfico 1, as publicações com links para documentos científicos representaram 24%.

Este resultado demonstra que existe um processo descentralizado de indexação proporcionado pela Folksonomia que cumpre "a principal finalidade de um sistema de recuperação da informação [que] é fornecer aos usuários a informação na forma e momento exigidos por eles" (CARNEIRO, 1985, p. 223).

Com essa premissa é possível observar que as redes sociais cumprem esse pré-requisito, pois permitem que informações sejam compartilhadas e recuperadas, seja no formato de um *tweet*, um post no Facebook ou uma imagem no Instagram. Essa diversidade de formatos segue em consonância com o pensamento de Gil Leiva e Fujita (2012) ao afirmarem que "qualquer objeto pode ser indexado, ou seja, reduzido a representações conceituais que facilitem seu armazenamento e recuperação em bases de dados" (GIL LEIVA; FUJITA, 2012, p. 65).

Além disso, também demonstra a possibilidade da utilização das hashtags para a identificação do tipo de produto científico compartilhado e a área do saber predominante, para cada tipo de área. Para os Estudos Métricos da Informação, esses indicadores podem auxiliar na construção e desenvolvimento de novos sistemas capazes de auxiliar instituições de pesquisas, universidades e pesquisadores a incorporar camadas mais refinadas aos dados altmétricos.

Ao analisar o instrumental de ferramentas utilizadas é possível observar que o TAGSExplorer, para a captura de tweets, é uma ferramenta que entregou volume de conteúdos superior ao esperado. No entanto, após ultrapassar o limite de 200 mil tweets começou a enviar mensagens de erro, com a informação que a coleta seria refeita. Esse tipo de erro pode ser prejudicial para grandes volumes de dados. No entanto, com essa observação, para trabalhos posteriores serão realizadas quebras temporais, para evitar que aconteça a sobrecarga da ferramenta.

Enquanto, a limpeza, modelagem e análise dos conteúdos com a utilização do software R, demonstrou grandes possibilidades para esse tipo de estudo. As bibliotecas e/ou pacotes auxiliaram na modelagem dos dados, principalmente, no caso desta pesquisa, na leitura de links encurtados.

4.1 Discussão teórica

A popularização dos meios digitais trouxe para o grande público a possibilidade de abandonar o papel de consumidor passivo de informações para ser o seu produtor. Essa nova realidade, diferente da percepção de Johnson (2003) que acreditou que esse novo ambiente acabaria por ser tornar um "caos informacional", na prática acabou por criar uma "indexação democrática", conforme postulado por Brown, Hidderley, Griffin e Rollason (1996), ao desenvolver um projeto para a indexação e recuperação de imagens, por pessoas comuns.

Essa democratização também foi identificada por Lévy (2010) ao definir a contribuição do público como "inteligência coletiva", pois existe uma contribuição

baseada em percepções e inteligência individual. A percepção é complementada por Gil Leiva (1997) ao destacar que a explosão informacional proporcionou que na prática a indexação também fosse realizada por computadores, em um processo parcialmente ou totalmente automatizado. Enquanto Santos (2019) afirma que a Folksonomia pode ser utilizada como um conceito que "pode contribuir para os estudos métricos da informação científica" (SANTOS, 2019, p. 2).

Ao mesmo tempo, com esses dados preliminares, é possível identificar como as áreas do saber estão envolvidas tanto com o movimento do Open Access, como na utilização dos ambientes digitais para o compartilhamento e troca de informações sobre conteúdos científicos.

No Apêncide A, Gráfico 2, é possível identificar pelas hashtags a predominância de duas áreas: Ciências Exatas e da Terra e Ciências Biológicas, no compartilhamento de conteúdos.

4.2 A Folksonomia como suporte a indexação tradicional

Como afirmou Bossy (1995, não paginado, tradução nossa) "a diversidade de novos padrões de comunicação na rede eletrônica às vezes rompe as fronteiras entre circulação formal e informal, entre atividades que ocorrem dentro e fora dos laboratórios". A premissa permite trazer para o debate a Folksonomia e a Altmetria como campos ainda em fase de amadurecimento, mas que carregam em suas bases a utilização de técnicas e modelos tradicionais, adaptados para o novo ambiente. Essa utilização acaba por não gerar rupturas ou grandes impactos na forma como a academia e os pesquisadores observam nos resultados apresentados.

Extrair informações quantificáveis das redes sociais, por meio das *tags* vinculadas aos conteúdos demonstra ser a base para os estudos sobre a Folksonomia. No entanto, observa-se que o modelo adequado para taxonomia própria de exploração científica das redes sociais deve estar aliado às suas dinâmicas, ao invés, de apenas tentar buscar uma aproximação ou uma adaptação dos modelos tradicionais para esse novo ambiente.

Essa nova visão sobre a Folksonomia de não se debruçar somente sobre as tags, mas se apropriar de outros elementos, como os links, favorece os estudos do campo e pode auxiliar no desenvolvimento dos EMI em ambientes sociais.

5. Considerações Finais

A pesquisa demonstrou possibilidades, pontos positivos e limitações ao combinar a Folksonomia para extração de indicadores Altmetricos. No entanto, também identificou uma forte convergência entre os saberes, que ainda demonstra ser negligenciada, principalmente, nos Estudos Métricos da Informação.

Com a visão da possibilidade de uma convergência, foi construído um modelo de Altmétrico, enriquecido pela Folksonomia, que permitiu a delimitação de termos capturados, mas com capacidades de ampliar para o desenvolvimento de índices, rankings de produtividade, de dispersão e apropriação de conteúdos científicos nas redes sociais. Ao mesmo tempo, este tipo de modelagem demonstra a necessidade do aprimoramento dos sistemas de recuperação de recursos informacionais, como forma de refinar a coleta dos descritores sociais.

Assim, embora os resultados tenham sido positivos para a realização deste estudo, que permitiu identificar a viabilidade operacional do modelo, enquanto se avaliava as ferramentas que podem ser utilizadas. Ainda é necessário o desenvolvimento de um modelo mais robusto e funcional, que possa operar automaticamente somente com a inserção de dados, que é a próxima etapa da pesquisa.

No entanto, essas primeiras observações demonstram a que este tipo de estudo se enquadra nos debates sobre Gestão da Informação nos ambientes digitais, na transparência esperada dos Indicadores Métricos e em novas metodologias para os Estudos Métricos da Informação.

6. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001 e

Conselho Nacional de Desenvolvimento Científico e Tecnológico, Processo 430982/2018-6 e 315521/2020-1.

Referências

BOSSY, M. J. The Last of the Litter: Netometrics. 1995. Disponível em: http://gabriel.gallezot.free.fr/Solaris/d02/2bossy.html. Acesso em: 25 ago. 2022.

BROWN, Pauline; HIDDERLEY, Rob; GRIFFIN, Hugh; ROLLASON, Sarah. The democratic indexing of images. **New Review of Hypermedia and Multimedia**, v. 2, n. 1, p. 107-120, 1996.

CARNEIRO, MARÍLIA VIDIGAL. Diretrizes para uma política de indexação. **Revista da Escola de Biblioteconomia da UFMG**, v. 14, n. 2, 1985.

DORA, S. F. San Francisco declaration on research assessment. 2012. Disponível em: https://sfdora.org/read/read-the-declaration-portugues-brasileiro/. Acesso em: 11 ago. 2021.

FURNER, Jonathan. Folksonomies. In: **Encyclopedia of Library and Information Sciences, Third Edition**. CRC Press, 2009. p. 1858-1866

GIL LEIVA, Isidoro; FUJITA, Mariângela Spotti Lopes (Ed.). **Política de indexação**. Editora Oficina Universitária, 2012.

GIL LEIVA, Isidoro. **La automatización de la indización, propuesta teórico-metodológica: aplicación al área de biblioteconomía y documentación**. Murcia, España: Universidad de Murcia, 1997.

GUEDES, R. de M.; DIAS, Eduardo José Wense. Indexação social: abordagem conceitual. **Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis**, v. 15, n. 1, p. 39-53, 2010.

HAUSTEIN, S.; BOWMAN, T. D.; COSTAS, R. Interpreting “altmetrics”: viewing acts on social media through the lens of citation and social theories. In: **Theories of Informetrics: A Festschrift in Honor of Blaise Cronin**. Berlin: Cassidy R. Sugimoto (Ed.), 2015. p. 372–405.

HAWKSEY, M. **TAGS**. , 2014. Disponível em: <https://tags.hawksey.info/>

HOFFMANN, C. P.; LUTZ, C.; MECKEL, M. A relational altmetric? Network centrality on ResearchGate as an indicator of scientific impact. **Journal of the Association for Information Science and Technology**, v. 67, n. 4, p. 765–775, 2016.

HOTHO, Andreas. Information retrieval in folksonomies: Search and ranking. In: **European semantic web conference**. Springer, Berlin, Heidelberg, 2006. p. 411-426.

JOHNSON, Steven. **Emergência: a dinâmica de rede em formigas, cérebros, cidades e softwares**. Editora Schwarcz-Companhia das Letras, 2003.

MAI, Jens-Erik. Folksonomies and the new order: authority in the digital disorder. **KO KNOWLEDGE ORGANIZATION**, v. 38, n. 2, p. 114-122, 2011.

LEVY, Pierre. **Cibercultura**. Editora 34, 2010.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up? Sentiment Classification using Machine Learning Techniques. *In*: PROCEEDINGS OF THE ACL-02 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING - EMNLP ’02 2002, Morristown, NJ, USA. **Anais** [...]. Morristown, NJ, USA: Association for Computational Linguistics, 2002. Disponível em: http://portal.acm.org/citation.cfm?doid=1118693.1118704. Acesso em: 18 ago. 2022.

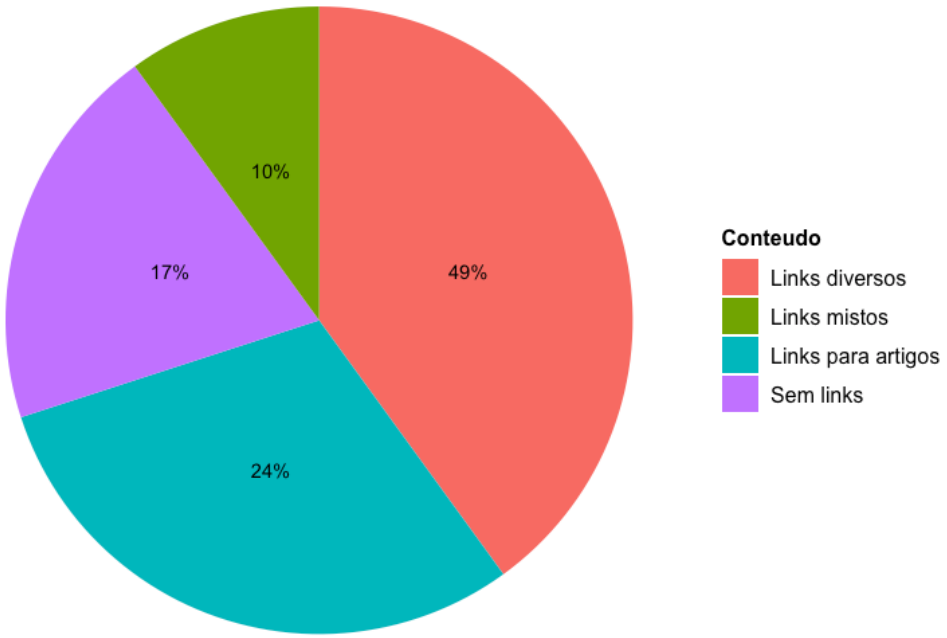
RSTUDIO TEAM. **RStudio: Integrated Development Environment for R**. : Prairie Trillium.Boston, MARStudio, PBC, , [s.d.]. Disponível em: <https://www.rstudio.com>

SANTOS, Raimunda Fernanda dos. A Folksonomia e o seu impacto na comunicação científica. **Revista Conhecimento em Ação**, v. 4, n. 2, p. 1-3, 2019.

ZOLLER, Daniel. Posted, visited, exported: Altmetrics in the social tagging system BibSonomy. **Journal of Informetrics**, v. 10, n. 3, p. 732-749, 2016.

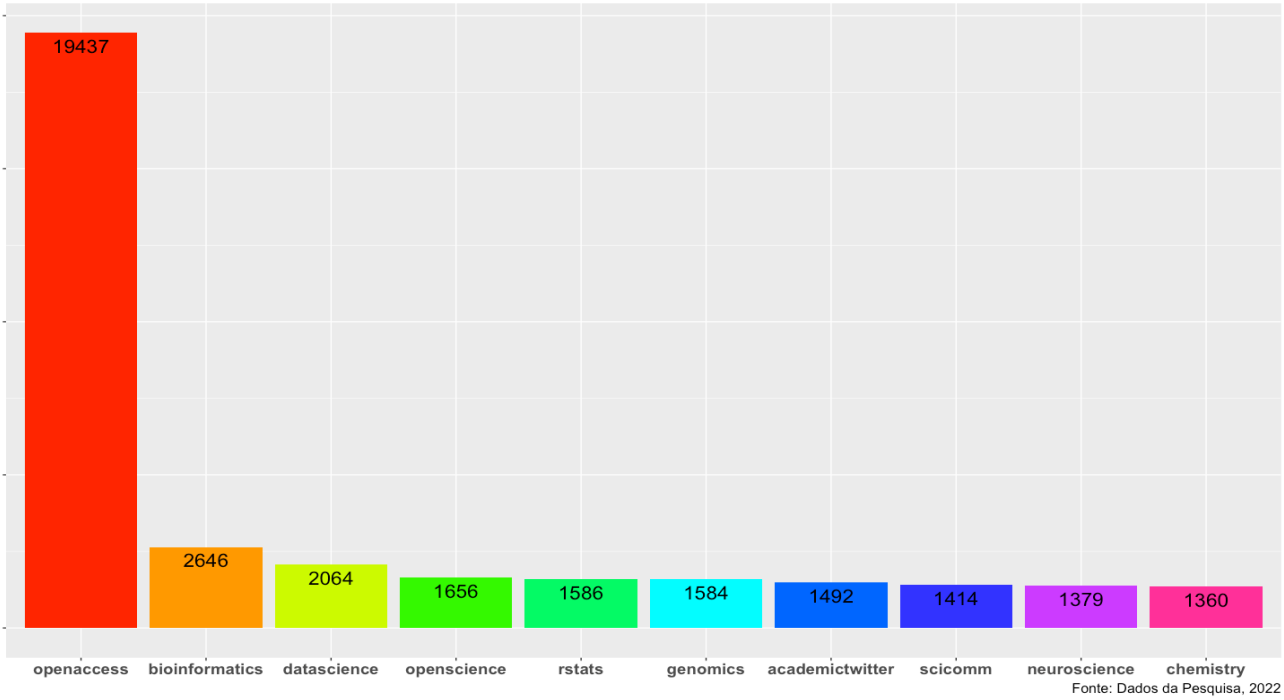
7. Apêndice A – Gráficos

Gráfico 1 - Perfil do conteúdo gerado pelo público do Twitter



Dados da Pesquisa, 2022

Gráfico 2 - Amostra dos resultados das hashtags com mais de 1000 menções



Fonte: Dados da Pesquisa, 2022

A INDÚSTRIA 4.0 SOB O PRISMA DA CIÊNCIA DE DADOS: UMA PROPOSTA DE MODELO DE MATURIDADE

INDUSTRY 4.0 UNDER THE PRISM OF DATA SCIENCE: A PROPOSAL FOR A MATURITY MODEL

Jacqueline Zonichenn Reis[1]
(1) Universidade Paulista, Rua Dr. Bacelar 1212 - São Paulo/SP, zonichenn@hotmail.com

Resumo
O objetivo deste artigo é propor um modelo de maturidade da Indústria 4.0 sob o prisma da ciência de dados. A Indústria 4.0 se tornou referência no desenvolvimento de aplicações e tecnologias para atender aos requisitos de uma produção mais ágil e inteligente. Entretanto, existem lacunas sobre quais tecnologias se adequam melhor à uma determinada cadeia produtiva ou quais seriam suficientes para classificá-la como já inserida na Indústria 4.0. Os modelos de maturidade existentes trazem fases deste avanço porém sem indicar como alcançá-las. A partir da revisão da literatura, observou-se dois fatores importantes para tal evolução: a conectividade; e a obtenção, análise e uso de dados. Neste contexto, um modelo teórico-conceitual foi proposto para traçar cenários da Indústria 4.0 com relação à adoção destas inovações. O modelo apresenta dois eixos, conectividade e ciência de dados, que resulta em quatro quadrantes ou cenários de maturidade: automação; digitalização; visibilidade e análise descritiva; e adaptabilidade e análise preditiva. Este trabalho, ainda preliminar, visa apoiar gestores e profissionais a identificar em que patamar estaria determinado negócio, bem como embasar os investimentos em tecnologia e os impactos esperados. O artigo também deriva recomendações para pesquisas futuras sobre o tema.

Palavras-chave: Indústria 4.0; Modelo de maturidade; Ciência de Dados.

Abstract
The purpose of this article is to propose an Industry 4.0 maturity model from the perspective of data science. Industry 4.0 has become a reference in the development of applications and technologies to meet the requirements for an agile and intelligent production. However, there are gaps about which technologies best suit a given production chain or which would be sufficient to classify it as Industry 4.0. Existing maturity models bring stages for advancing but without indicating how to achieve them. From the literature review, two important factors for such evolution were observed: the connectivity; and the collection, analysis, and use of data. In this context, a theoretical-conceptual model was proposed to outline Industry 4.0 scenarios regarding the adoption of these innovations. The model presents two axes, connectivity, and data science, which results in four quadrants or maturity scenarios: automation; digitization; visibility and descriptive analysis; adaptability and predictive analytics. This study, which is still preliminary, aims to support managers and professionals in identifying the level at which a given business would be, as well as providing a basis for investments in technology and the expected impacts. The article also derives recommendations for future research on the topic.

Keywords: Industry 4.0; Maturity model; Data Science.

1 Introdução

Em um mercado dinâmico e competitivo, a indústria vem adaptando suas operações para atender aos requisitos de uma produção cada vez mais ágil e inteligente. O período denominado Quarta Revolução Industrial, ou Indústria 4.0, se torna referência na aplicação de inovações e tecnologias na manufatura. Destacam-se tecnologias que buscam integrar internet das coisas e de serviços, sistemas cyberfísicos, big data, computação na nuvem, inteligência artificial, realidade aumentada, manufatura aditiva, entre outros (ZUTIN et al., 2022)

Diferentes abordagens são sugeridas como caminho a ser seguido para o alcance da Indústria 4.0. Porém, surgem lacunas em se determinar qual conceito ou tecnologia devem ser adotados (ALCÁCER; CRUZ-MACHADO, 2019). A mesma dificuldade é percebida com relação ao que seria pré-requisito para classificar uma determinada fábrica como já inserida na indústria 4.0; ou ainda operando com a mesma automação, algo que seria fruto, na verdade, da revolução industrial anterior. Jeremy Rifkin sugere, inclusive, que muitas das inovações que dizem pertencer à quarta revolução industrial são oriundas do surgimento da computação desde os anos 1960 como parte da terceira revolução industrial (PRISECARU, 2016).

Pesquisas recentes indicam que a transição para a Indústria 4.0 não se traduz em realidade para a maioria das empresas. De acordo com Ruggero et al. (2020), falta

infraestrutura, e a adaptabilidade ainda não é tecnicamente possível. Torna-se relevante, portanto, identificar não só o que as empresas devem tratar como pré-requisito para a maturidade da Indústria 4.0, mas também como acessar em que estágio elas estão, para que assim possam melhor direcionar as ações e os investimentos de forma assertiva.

Na literatura científica e empírica sobre a Indústria 4.0, existem modelos de maturidade que mostram fases deste processo e servem como uma orientação para as empresas. Porém, estes modelos geralmente abordam níveis de maturidade em camadas, como uma forma linear, em termos de conexão com a internet das coisas e digitalização. Este trabalho traz um enfoque voltado para a ciência de dados, além dos domínios de infraestrutura e conectividade.

O artigo propõe um modelo teórico-conceitual, em que a ciência de dados componha o quadro de maturidade da Indústria, trazendo uma contribuição para a discussão sobre o tema. Os resultados trazem um modelo de distribuição em dois eixos: conectividade e ciência de dados. A menor ou maior adesão a essas tecnologias resulta em quatro quadrantes ou cenários de maturidade. No primeiro quadrante estaria a automação, seguida pela digitalização, no segundo. No terceiro estariam a visibilidade e a análise descritiva, até se chegar no quarto quadrante em que a Indústria 4.0 atinge sua maturidade, através da adaptabilidade e análise preditiva.

2 Fundamentação teórica

Alguns conceitos serão explorados nesta seção para melhor entendimento do artigo.

2.1 Da primeira à quarta revolução Industrial

As três primeiras revoluções industriais trouxeram mudanças significativas para o processo de manufatura; desde os motores a vapor até a produção com o uso da energia elétrica e, posteriormente, a automação e computação (WAHLSTER, 2012). Essa evolução se deu devido à mecanização, eletricidade e a tecnologia da informação, respectivamente. A Tabela 1 mostra as contribuições das revoluções industriais.

Tabela 1 – Revoluções Industriais (adaptado de Prisecaru, 2016)

Revoluções	Período	Principais inovações
1ª Rev Industrial	1760-1850	Máquina a vapor
2ª Rev Industrial	1850-1945	Motor de combustão interna
3ª Rev Industrial	1950-2010	Computadores Robôs
4ª Rev Industrial	a partir de 2011	Internet, Impressão 3D Engenharia genética

Na segunda década do século XXI, inicia-se o período considerado Quarta Revolução Industrial, ou Indústria 4.0. O termo Indústria 4.0 foi enraizado na estratégia do governo alemão em 2011, com o objetivo de ganhar força na manufatura global por meio da aplicação avançada de sistemas de informação e comunicação. Através detas novas tecnologias, o ambiente fabril se torna inteligente e possibilita a customização em massa (KAGERMANN; WAHLSTER; HELBIG, 2013).

Entre as tecnologias, as principais seriam a Internet das Coisas (IoT, ou *Internet of Things*), Internet de Serviços (IoS, ou *Internet of Services*), e os Sistemas CiberFísicos (CPS, ou *CyberPhysical Systems)*. Os CPS são sensores e atuadores que monitoram processos físicos e criam uma cópia virtual do mundo físico (JAZDI, 2014). Pela IoT, os CPS comunicam e cooperam entre si e com humanos em tempo real (GIVEHCHI et al., 2017). Em seguida, através da IoS, serviços são oferecidos na nuvem e utilizados pelos participantes da cadeia de valor. A IoS seria uma Internet do Futuro que detecta e usa informações contextuais para se adaptar a um cenário imprevisível, permitindo a configuração *ad-hoc* de novos modelos de negócios (REIS et al., 2022).

2.3 Ciência de dados

A ciência de dados é definida como a extração de conhecimento de dados de alto volume, usando habilidades em ciência da computação, estatística e o conhecimento de especialistas (NASUTION, 2021).

O desenvolvimento da ciência de dados tem sido motivado pela explosão de dados neste momento de transformação digital. Vicario e Coleman (2020) ressaltam que muitas das técnicas matemáticas, estatísticas

e de aprendizado de máquina que envolvem a ciência de dados já existem há muitos anos. O que mudou foi a disponibilidade de maiores quantidades de dados coletados através da IoT e que agora podem ser armazenados ao invés de serem meramente observados e sobrescritos. Isso se deve a um conjunto de fatores, como maior acesso à Internet, barateamento de sensores, maior capacidade e disponibilidade de recursos computacionais, entre outros (APPIAH-OTOO; SONG, 2021).

Atualmente, as técnicas de ciência de dados têm aplicações em quase todos os campos, e estão sendo aplicadas também à Indústria 4.0. Nesta configuração avançada, dados massivos são criados e armazenados a cada segundo. Especialistas com experiência em matemática e computação avançada são requisitados para identificar as causas-raiz de falhas e desvios de qualidade das máquinas. Além disso, novos elementos com propriedades personalizadas podem ser descobertos com teorias de materiais e habilidades computacionais. A integração da ciência de dados com a indústria 4.0 aumentará a eficiência e será útil para promover a qualidade do material, contribuindo para minimizar a perda de tempo e dinheiro (SAJID et al., 2021).

3 Procedimentos Metodológicos

O método utilizado para escrever o artigo foi a pesquisa exploratória. Através de uma revisão bibliográfica, não sistemática, foram encontradas referências de modelos de maturidade que a indústria já utiliza. Em seguida, fazendo um contraponto a estes modelos, foi proposto um novo modelo teórico-conceitual.

3.1 Modelos de maturidade da Indústria 4.0

Um modelo conhecido para implantação da Indústria 4.0 é o da Acatech (SCHUH et al., 2020), representado pela Figura 1. São consideradas seis etapas, sendo as duas primeiras a informatização e a conectividade, que fazem parte da fase de digitalização. Nessa fase são estabelecidas as conexões entre os dispositivos, sensores, máquinas e software, possibilitando a comunicação do mundo físico com o virtual. Em seguida, na fase da Indústria 4.0 propriamente dita, estão as demais etapas: visibilidade, transparência, capacidade preditiva e adaptabilidade.

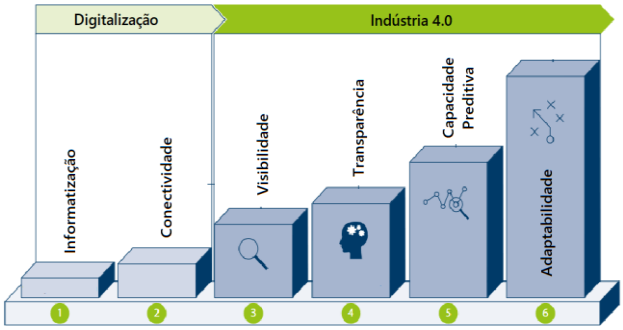


Figura 1 – Modelo Acatech de maturidade da Indústria 4.0 (adaptado de SCHUH et al., 2020)

3.2 Modelo de maturidade proposto

É crescente o número de publicações na literatura científica que buscam ou propõe novos modelos para a avaliação do nível de maturidade da Indústria 4.0. Tais modelos sempre variam de 4 a 6 níveis, sendo que cada nível seria um degrau nesta evolução (SILVA; BARBALHO, 2019).

Um aspecto importante da Indústria 4.0 é a obtenção de dados em tempo real e seu tratamento para obtenção de informações atuais e precisas que possibilitam melhores decisões. A troca ágil de informações entre subsistemas permite a colaboração para a atuação em questões que afetam a cadeia de produção como um todo. Destacam-se dessa forma dois fatores importantes para tal integração: a conectividade para troca de dados; e a obtenção, análise e uso dos dados (SANTOS; RUGGERO; SILVA, 2021).

Ao longo da implementação da Indústria 4.0, muitas empresas tendem a investir mais em conectividade do que em ciência de dados. O inverso também pode acontecer e, baseado nesta dicotomia, propomos um modelo distribuído em dois eixos, ao invés de camadas como os outros. Na figura 3 é apresentado o modelo proposto.

4 Resultados e discussão

O modelo proposto na Figura 3 mostra as diferentes fases da Indústria com relação à ciência de dados (eixo x) e conectividade (eixo y). Quatro cenários ou quadrantes se formam a partir desta distribuição.

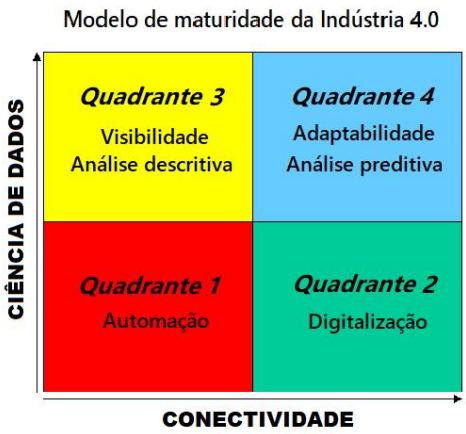


Figura 3 – Modelo de maturidade da Indústria 4.0 (da autora)

4.1 Cenários de maturidade

- Automação (quadrante 1 ou Q1): Estágio inicial em que a produção é automatizada, principalmente nas tarefas manuais e repetitivas. A automação é algo que já se inicia na Terceira Revolução Industrial, por isso já existem a informatização e a robotização. Entretanto, a conectividade e a ciência de dados ainda não são bem trabalhadas neste cenário.

- Digitalização (quadrante 2 ou Q2): Neste estágio ocorre uma transposição de processos, serviços ou produtos para o meio digital. Os sistemas CiberFísicos criam uma cópia virtual do mundo físico. Ocorre um investimento maior em conectividade, porém isso ainda não seria suficiente para se ter a Indústria 4.0 em sua plenitude. A digitalização é vista, de forma errônea, como uma transformação digital, porém essa última implica em mudanças muito mais profundas no negócio da organização, impactando sua estrutura e sua estratégia competitiva. A digitalização traz melhorias, mas sem impacto nestes aspectos.

- Visibilidade e Análise Descritiva (quadrante 3 ou Q3): Estágio em que se torna possível visualizar o andamento das operações, baseando-se na coleta e análise dos dados de diferentes processos. Consegue-se a visibilidade e análise descritiva, que seria a análise voltada para o passado, proporcionando uma melhora na tomada de decisão. O que predomina é um investimento em ciência de dados, maior do que em conectividade. Fazendo um paralelo com Q2, neste estágio de Q3 os processos são melhor estudados para que os dados sejam de qualidade, e para que seja feita também uma análise destes dados, ao invés de somente uma coleta dos mesmos de forma aleatória.

- Adaptabilidade e Análise Preditiva: (quadrante 4 ou Q4): Neste estágio tem-se o funcionamento pleno da Indústria 4.0, que envolve não apenas uma rede conectada e a troca automática de informações, mas também um sistema que se tornou consciente e inteligente o suficiente para prever e manter máquinas, controlar o processo de produção, e gerenciar o sistema fabril de forma descentralizada. Através dos sistemas de informação integrados entre a produção e os ERP, consegue-se a tomada de decisão em tempo real para se reagir às mudanças do ambiente de forma contextual e adocrática, reconfigurando unidades e retro-alimentando os sistemas.

4.2 Validação dos cenários

O presente trabalho é preliminar e deve ser validado em estudos de casos ou pesquisas futuras.

O modelo proposto pode ser ilustrado considerando o seguinte cenário: uma indústria siderúrgica utiliza um sistema de filtros de emissão, fornecido por um parceiro. Sua capacidade produtiva é volátil e exige elasticidade, bem como taxas de repasse e atualização das regras de alocação. Quando há um aumento na demanda de produção, a capacidade de filtragem deve ser maior para atender o novo volume de emissão. Por outro lado, quando a demanda cai, a indústria não deve pagar por uma filtragem de emissão mais alta. Porém é isso que acontece em um cenário de automação, do Quadrante 1 do modelo proposto.

Mesmo que sensores sejam instalados e permitam uma conectividade dos filtros, o sistema se torna digital, porém ainda não se

torna inteligente. Sensores que alertam quando aumentar ou diminuir a capacidade manualmente, estão digitalizando parte do processo, mas não o tornam mais ágil. Esse seria o cenário de digitalização do quadrante 2 do modelo.

Aplicando a ciência de dados e coletando os dados fornecidos sobre o uso dos filtros, obtem-se mais visibilidade. Análises desses dados ajudam a prever uma sazonalidade e aumentam a capacidade descritiva para uma tomada de decisões futuras, o que configura o cenário do quadrante 3 do modelo.

Se a análise dos dados e o controle de acesso ocorrem em tempo real por sistemas de informação, a precisão da demanda se torna maior. Os sensores conectados com equipamentos "inteligentes" possibilitam a tomada de decisão descentralizada e em tempo real. Além disso, contratos inteligentes podem reduzir a burocracia e automatizar as regras contratuais, substituindo os contratos de serviços convencionais. Esse modelo orientado ou "pay-per-use" é um exemplo do cenário de adaptabilidade do quadrante 4 do modelo.

4.3 Discussão dos resultados

A automação e digitalização, quadrantes Q1 e Q2 respectivamente, podem sim proporcionar a abertura de novos mercados, alavancar a inovação e os ganhos de produtividade, especialmente nas economias em desenvolvimento. Porém isso ainda não é a Indústria 4.0. Como o Jeremy Rifkin sugere, muitas das inovações digitais que dizem pertencer à quarta revolução industrial devem ser agrupadas com a computação e acesso à Internet oriundos da terceira revolução industrial (PRISECARU, 2016).

Ao se investir mais em ciência de dados consegue-se deslocar para os quadrantes Q3 e Q4. Enquanto a ciência de dados trabalha no acesso, processamento, análise dos dados, e soluções para tomadas de decisão, utilizando uma abordagem "top-down" que seria orientada ao negócio; a conectividade foca nos dispositivos e na conexão dos mesmos aos sistemas.

Em Q4 é que a real convergência entre Tecnologia de Informação (IT ou *Information Technology*) e Tecnologia de Operações (OT ou *Operational Technology*) acontece. Com estudo e investimento adequado, não só em conectividade, mas também em ciência de dados, consegue-se a adaptabilidade e análise preditiva que seriam as premissas para se ter uma Indústria 4.0 mais madura.

Enquanto os CPS e a IoT permitem a aquisição de dados, como por exemplo o consumo de energia e a análise no nível das máquinas e da linha de produção, os sistemas inteligentes fazem o processamento e análise desse volume massivo de dados em tempo real, conectando a cadeia de valor e gerando novos serviços e negócios.

5 Considerações Finais

O objetivo deste artigo foi propor um modelo de maturidade da Indústria 4.0 sob a perspectiva da ciência de dados. A partir de uma revisão bibliográfica, são trazidos conceitos sobre o tema e um modelo teórico é proposto.

O modelo resulta em quatro quadrantes que caracterizam a Indústria 4.0 com relação à uma menor ou maior aplicação da ciência de dados. Os quadrantes são: automação; digitalização; visibilidade e análise descritiva; adaptabilidade e análise preditiva. Com o modelo distribuído em quadrantes pelos eixos de conectividade e ciência de dados, fica mais claro que a coleta de dados e a oportunidade de analisá-los para realizar processos mais avançados são cruciais para a maturidade da Indústria 4.0. O modelo pode servir como referência aos profissionais quando forem avaliar o nível de maturidade de determinado negócio. A pesquisa pode ainda derivar estudos de caso dentro destes cenários, o que é uma oportunidade para trabalhos futuros.

O trabalho mostra a relevância da ciência de dados na distribuição de tecnologia e investimento para que seja alcançado um patamar diferenciado na Indústria 4.0, revertendo este ganho para os gestores, acionistas, governo e sociedade.

Referências

ALCÁCER, V.; CRUZ-MACHADO, V. Scanning the Industry 4.0: A Literature Review on Technologies for Manufacturing Systems. **Engineering Science and Technology, an International Journal**, v. 22, n. 3, p. 899–919, jun. 2019.

APPIAH-OTOO, I.; SONG, N. The impact of ICT on economic growth-Comparing rich and poor countries. **Telecommunications Policy**, v. 45, n. 2, p. 102082, mar. 2021.

GIVEHCHI, O. et al. Interoperability for Industrial Cyber-Physical Systems: An Approach for Legacy Systems. **IEEE Transactions on Industrial Informatics**, v. 13, n. 6, p. 3370–3378, dez. 2017.

JAZDI, N. **Cyber physical systems in the context of Industry 4.0**. 2014 IEEE International Conference on Automation, Quality and Testing, Robotics. **Anais**... Em: 2014 IEEE INTERNATIONAL CONFERENCE ON AUTOMATION, QUALITY AND TESTING, ROBOTICS (AQTR). Cluj-Napoca, Romania: IEEE, maio 2014.

KAGERMANN, H.; WAHLSTER, W.; HELBIG, J. **Securing the future of German manufacturing industry: Recommendations for implementing the strategic initiative INDUSTRIE 4.0 – Final Report of the Industrie 4.0 working group. Acatech – National Academy of Science and Engineering**. [s.l: s.n.].

NASUTION, M. K. M. Industry 4.0: Data science perspective. **IOP Conference Series: Materials Science and Engineering**, v. 1122, n. 1, p. 012037, 1 mar. 2021.

PRISECARU, P. Challenges of the fourth industrial revolution. **Knowledge Horizons. Economics**, v. 8, n. 1, p. 57–62, 2016.

REIS, J. Z. et al. Business Models for the Internet of Services: State of the Art and Research Agenda. **Future Internet**, v. 14, n. 3, p. 74, 25 fev. 2022.

RUGGERO, S. M. et al. Industry 4.0: Maturity of Automotive Companies in Brazil for the Digitization of Processes. Em: LALIC, B. et al. (Eds.). **Advances in Production Management Systems. The Path to Digital Transformation and Innovation of Production Management Systems**. IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing, 2020. v. 591p. 131–138.

SAJID, S. et al. Data science applications for predictive maintenance and materials science in context to Industry 4.0. **Materials Today: Proceedings**, v. 45, p. 4898–4905, 2021.

SANTOS, N. A. DOS; RUGGERO, S. M.; SILVA, M. T. DA. Indústria 4.0 no Brasil: desafios do segmento automotivo para integração da cadeia de suprimentos. **Research, Society and Development**, v. 10, n. 8, p. e18110817251, 10 jul. 2021.

SCHUH, G. et al. **Industrie 4.0 Maturity Index. Managing the Digital Transformation of Companies – UPDATE 2020**. [s.l.] Acatech – National Academy of Science and Engineering, 22 abr. 2020.

SILVA, I. A. DA; BARBALHO, S. C. M. **MODELOS DE MATURIDADE: DO CMM AOS MODELOS PARA INDÚSTRIA 4.0**. Blucher Engineering Proceedings. **Anais**... Em: 12º CONGRESSO BRASILEIRO DE INOVAÇÃO E GESTÃO DE DESENVOLVIMENTO DE PRODUTO. São Paulo: Editora Blucher, nov. 2019.

VICARIO, G.; COLEMAN, S. A review of data science in business and industry and a future view. **Applied Stochastic Models in Business and Industry**, v. 36, n. 1, p. 6–18, jan. 2020.

WAHLSTER, W. From Industry 1.0 to Industry 4.0: Towards the 4th Industrial Revolution, Forum Business meets Research. Em: [s.l: s.n.].

ZUTIN, G. C. et al. Readiness levels of Industry 4.0 technologies applied to aircraft manufacturing—a review, challenges and trends. **The International Journal of Advanced Manufacturing Technology**, v. 120, n. 1–2, p. 927–943, maio 2022.

A RELAÇÃO DA ORGANIZAÇÃO DO CONHECIMENTO COM AS HUMANIDADES DIGITAIS A PARTIR DAS DEZ PREMISSAS DE BARITE

THE RELATIONSHIP OF KNOWLEDGE ORGANIZATION WITH DIGITAL HUMANITIES FROM THE TEN PREMISES OF BARITE

Ana Cristina de Albuquerque[1], Marcos Antonio de Moraes[2] ,Ania Rosa Hernandez Quintana[3]
(1) Universidade Estadual de Londrina (UEL), PPGCI UEL, Londrina, Brasil, albuanati@uel.br
(2) Universidade Estadual de Londrina (UEL), Londrina, Brasil, marcosmoraes@uel.br
(3) Universidad de La Habana, PPGCI, La Habana, Cuba, aniahernandez.quintana@gmail.com

Resumo
Questiona como pode ser problematizada a relação da Organização do Conhecimento com as Humanidades Digitais a partir de bases teóricas da OC e das HDs. O estudo é de base bibliográfica e exploratória. Como percurso metodológico, utilizou-se a Análise de Conteúdo, considerando-se as dez premissas da Organização do Conhecimento apresentadas por Barité (2001). Foram elencadas cinco características das Humanidades Digitais e, a partir das dez premissas, discutidas suas interlocuções. Refletir sobre a relação da Organização e Representação do Conhecimento com as Humanidades Digitais a partir das 10 (dez) premissas pode ajudar a pensar as formas de colaboração das áreas em relação a disponibilização do conhecimento de forma crítica e ética. Conclui que as premissas da OC, combinadas com as características das HDs, têm o potencial de elaboração de um controle nas relações conceituais, além de materializar alguns objetivos das HDs, como o de disponibilizar e disseminar o conhecimento por meio digital, fornecer bases integradas às Ciências Humanas e Sociais no sentido de metodologias e teorias e potencializar as atividades digitais diante das perspectivas da humanização visando os usuários, seus relacionamentos com os conceitos e contextualização dos documentos.

Palavras-chave: Organização do Conhecimento; Humanidades Digitais; Representação do Conhecimento.

Abstract
It questions how the relationship between the Knowledge Organization and the Digital Humanities can be problematized from theoretical bases of OC and HDs. The study is bibliographic and exploratory. As a methodological approach, Content Analysis was used, considering the ten premises of the Knowledge Organization presented by Barité (2001). Five characteristics of the Digital Humanities were listed and, based on the ten premises, their interlocutions were discussed. Reflecting on the relationship of the Organization and Representation of Knowledge with the Digital Humanities from the 10 (ten) premises can help to think about the ways of collaboration of the areas in relation to the availability of knowledge in a critical and ethical way. It concludes that the assumptions of the OC, combined with the characteristics of the HDs, have the potential to elaborate a control in the conceptual relationships, in addition to materializing some objectives of the HDs, such as making knowledge available and disseminating digitally, providing integrated bases to the Human and Social Sciences in the sense of methodologies and theories and to enhance digital activities in the face of the perspectives of humanization aimed at users, their relationships with the concepts and contextualization of documents.

Keywords: Knowledge Organization; Digital Humanities; Knowledge Representation.

1 Introdução

Com aporte de fundamentos científicos e princípios teóricos-metodológicos, incursões são realizadas quanto ao processo de sistematização da organização, representação e recuperação do conhecimento, que, consequentemente, será representado, respeitando-se o contexto e as necessidades dos usuários que virão a acessar os ambientes em que são disseminados. Assim, a organização e representação do conhecimento que é gerado de maneira planejada ou não, mas que, quando institucionalizado a partir dos suportes informacionais, necessita de tratamento para que as informações contidas nos documentos possam ser recuperadas, volta-se para as representações e ordenamento conceitual do conhecimento que está histórica e epistemologicamente vinculado ao estudo das estruturas conceituais que, de acordo com Dahlberg (1993), é uma unidade do conhecimento,

representado, entre outras formas, através das classificações.

Segundo Barité (2001, p. 39-40): a Organização do Conhecimento (OC): "[...]procura, entonces, brindar um continente conceptual adecuado a las diversas prácticas e actividades sociales vinculadas con el acceso al conocimiento [...]", como o tratamento e gestão de uso da informação. O autor destaca que a OC se caracteriza também como agregadora dos fenômenos teóricos e das atividades aplicadas que, encadeadas, vinculam a estruturação, disposição, acesso e a difusão do conhecimento socializado. Portanto, os aspectos característicos da OC estão ligados ao compartilhamento e melhores formas de atingir o usuário, por meio da recuperação de informações que dialoguem com suas escolhas e contextos.

Por sua vez, as Humanidades Digitais (HDs), podem ser descritas como diferentes práticas que convergem interdisciplinarmente no âmbito das aplicações das tecnologias computacionais. Nas Ciências Humanas e Sociais, Artes e Letras, as HDs movimentam instrumentos e perspectivas referentes ao mundo digital, formando comunidades de práticas que têm a finalidade de contribuir com o avanço, disseminação e compartilhamento do conhecimento, com foco na qualidade das pesquisas que podem ser realizadas e procurando enriquecer os saberes coletivos (MANIFESTO DE HUMANIDADES DIGITAIS, 2011).

A relação entre HDs e Ciência da Informação vem sendo trabalhada por autores brasileiros, como Pimenta (2016, 2020), Castro e Pimenta (2018), Paletta (2018), que propõem discussões e ressaltam as especificidades de cada área, mas que sugerem uma aproximação ou possibilidades de interesses que sejam comuns, tendo como mediador o uso das tecnologias. No âmbito da OC, essa relação é demonstrada em estudos como o de Andrade e Dal'Evedove (2021), que realizaram um levantamento e análise de artigos que discutem as prováveis aproximações com o objetivo de: "[...] compreender como as teorias e tendências de se pensar as tecnologias digitais nos fazeres humanos são dialogadas na área." (ANDRADE e DAL'EVEDOVE, 2021). As autoras consideram que há estudos com debates que acolhem terminologicamente as questões referentes às HDs de forma a enfatizar os aspectos aplicados, mas que, discutir as teorias comuns ainda são escassas.

Em Albuquerque e Hernandez Quintana (2021), é realizado um debate sobre a interlocução dos Sistemas de Organização do Conhecimento e as Humanidades Digitais, a partir da abordagem da Análise de Domínio. As autoras demonstram elementos teóricos que se aproximam, mas que carecem de estudos mais aprofundados e que problematizem, por exemplo, como o aumento dos espaços informacionais e de acervos digitais podem otimizar a relação do usuário com o acesso às informações que correspondam ao seu universo ou como as questões referentes aos domínios do conhecimento e das comunidades discursivas serão representadas, a partir da interrelação dos produtos resultantes das HDs com os processos de tratamento informacional, que garantam as dimensões social, histórica e cultural da informação (HJØRLAND; ALBRETCHSEN, 1995). Nesse sentido, o presente trabalho questiona como pode ser problematizada a relação da Organização do Conhecimento com as Humanidades Digitais a partir de bases teóricas da OC e das HDs.

2 Objetivos

O objetivo do presente trabalho é refletir sobre a relação da Organização e Representação do Conhecimento com as Humanidades Digitais a partir das 10 (dez) premissas apresentadas por Barité (2011), premissas estas que, de acordo com o autor, possibilitam compreender o motivo de ser e justificar intelectualmente a Organização do Conhecimento.

3 Procedimentos Metodológicos

O estudo é de base bibliográfica e exploratório. Foi realizada a sistematização da literatura referente a Organização do Conhecimento, com ênfase nas dez premissas apresentadas por Barité (2001) e HDs. Como percurso metodológico, utilizou-se a Análise de Conteúdo, considerando-se as dez premissas referenciadas.

Os elementos de interlocução foram apresentados com inferências, a partir do

enfoque específico das dez premissas da OC, que perpassam por representar e organizar o conhecimento em diferentes domínios (BARITÉ, 2011). Os elementos das HDs foram baseados em Galina Russel (2011), Koh (2015), Pimenta (2016), Medeiros et al (2017), Paletta (2018) e Pimenta (2020), que conceituam e determinam as HDs em diálogo com a Ciência da Informação. Foram elencadas cinco características das Humanidades Digitais e, a partir das dez premissas, discutidas suas interlocuções.

4 Resultados

No capítulo intitulado "*Organización del Conocimiento: un nuevo marco teórico-conceptual en bibliotecologia y documentación*", Barité (2001), apresenta considerações sobre sobre como se define Organizaçao do Conhecimento e como pode se constituir um marco teórico conceitual para o campo científico.

O autor posiciona a OC enfatizando os aspectos sociais da informação e do conhecimento, e explica que está direcionada a fornecer subsídios teóricos e também se retroalimentar com os fenômenos que se relacionam, especificamente, como tratamento temático da informação e de forma mais ampla, como a gestão social da informação (BARITE, 2001).

A OC, associa aspectos teóricos e metodológicos de diferentes áreas do conhecimento, como a Linguística, Documentação, Informática, Filosofia, História da Ciência e Ciências Cognitivas, ressaltando os pontos interdisciplinares de forma mais fluída, possibilitando, desta forma, que constituam um terrreno fértil de contribuições e trocas que constroem um espaço formativo e de atuação que abre dialogos, preservando as especificidades de cada disciplina (BARITÉ, 2001). À vista disso, o autor define o objeto da Organização do Conhecimento como o conhecimento socializado, que proporciona pensar em técnicas para o desenvolvimento de Sistemas de Organização do Conhecimento e metodologias integradas tanto às perspectivas filosóficas, quanto à organização dos documentos em unidades de informação, dando espaço para elaborações interdisciplinares (BARITÉ, 2001).

Neste sentido, Barité (2001), propõe as 10 (dez) premissas com a justificativa de que as teorias buscam estabelecer um sistema coerente com concepções macro e que a explicação de fenômenos a partir destes sistemas, proporcionam determinar critérios, procedimentos, técnicas e princípios que contribuem para a consolidação de uma ciência em seus marcos teóricos e conceituais.

As dez premissas propostas, são as seguintes: 1. Conhecimento como produto, necessidade e dínamo social; 2. Conhecimento se dá a partir da informação e quando socializado, volta a ser informação; 3. A estrurura e a comunicação do conhecimento formam um sistema aberto; 4. O conhecimento deve ser organizado para seu melhor aproveitamento individual e social; 5. Existem várias formas possíveis de se organizar o conhecimento; 6. Toda organização do conhecimento é provisória, artificial e determinista; 7. O conhecimento é sempre registrado em documentos, se mostra como conjunto de dados organizados e permite usos indiscriminados; 8. O conhecimento se expressa em conceitos e se organiza a partir de sistemas de conceitos; 9. Os sistemas de conceitos são organizados para fins científicos, sistemáticos ou de documentação; 10. As leis que regem a organização de sistemas de conceitos são uniformes e previsíveis e se aplicam igualmente a qualquer área disciplinar.

A questão central disposta pelo autor é o conhecimento como produto e necessidade social, portanto socializado, que obedece a uma série de princípios que permitem sua organização, disseminação e recuperação. Barité (2001), transita entre o intelectual e o aplicado de maneira a demonstrar a organização do conhecimento com os Sistemas de Conceitos, sua temporalidade determinada e a hierarquia nas formas de organização.

Por sua vez, as HDs podem ser compreendidas como uma comunidade de práticas ou práticas mediadas pelas tecnologias de informação. De acordo com Hockey (2004), as HDs se caracterizam como uma nova área acadêmica fortemente

interdisciplinar, que é capaz de oferecer metodologias específicas oriundas das tecnologias digitais, que podem ser trabalhadas nos mais variados aspectos de investigação nas Humanidades. Moura (2019), explica que as HDs, enquanto campo, se tornou viável por conta da excessiva carga de digitalização de dados, que favoreceu uma expansão da produção científica, no sentido de maior relacionamento colaborativo e a distância, fato que altera significativamente as relações e as mídias antes utilizadas de forma analógica.

O termo Humanidades Digitais, segundo Rodríguez-Yunta (2014) e Galina Russel (2011), apresenta um crescimento de uso em todo o mundo, especialmente porque delimita o aspecto interdisciplinar que demonstra os processos relacionados ao uso de tecnologias digitais no âmbito das humanidades e, consequentemente, incide sob a demanda acadêmica, cultural e social de acessibilidade de fontes ligadas às humanidades que merecem reflexão, pois, permitem a discussão em torno da humanização das tecnologias, por meio dos trabalhos críticos realizados a partir da disponibilização dessas fontes.

Considerando as Humanidades Digitais não como algo apartado, mas sim, com uma continuidade dos processos de desenvolvimento de unidades informacionais, Almeida e Damian (2015), expõe que provavelmente pesquisadores, trabalham com as HDs, mas que não utilizam a denominação de tal. Os projetos, bancos de dados, bibliotecas digitais, repositórios, curadoria digital, incidem em um terreno, que as Humanidades Digitais se colocam e o esforço da percepção de que, a transferência de suportes e ambiências vão além, ou seja, vão em direção de métodos articulados.

Os estudos de Pimenta (2016), demonstram uma preocupação em definir aspectos teóricos e metodológicos das Humanidades Digitais. O autor explica que o campo se configura como espaço de estudo e pesquisa de ensino e de forma predominante como acesso à informação e inovação, se caracterizando assim como um campo híbrido (PIMENTA, 2016).

Neste sentido, Paletta (2018), define Humanidades Digitais como uma área dedicada às atividades acadêmicas entre as tecnologias digitais e humanidades. O autor explica que as HDs, por utilizarem aplicações e técnicas vinculadas a estes campos, podem propiciar tipos diversos de ensino e pesquisa, delineando um caminho de pensar a relação dos humanísticos com o digital e a disponibilização do patrimônio cultural.

Os objetivos das HDs, de acordo com Galina Russel (2011), versam sobre a criação de bases de dados digitais que contenham sistemas de recuperação, preservação, armazenamento da documentação e disseminação das informações dispostas; desenvolvimento de estruturas metodológicas que permitam criar novos documentos e dados sobre a documentação armazenada e constituir incentivo às pesquisas que elucidem o entendimento do novo campo.

Desta forma, de acordo com autores como Galina Russel (2011), Pimenta (2016), Koh (2015), Paletta (2018) e Medeiros et al (2017), os principais elementos característicos das Humanidades Digitais podem ser apontados como: ferramentas de aprendizagem, por permitir interação entre usuário e ambiente; possibilidade de acesso a grande quantidade de dados, sendo essencialmente analítica; têm o caráter colaborativo, por apresentarem documentos digitalizados e bancos de dados interligados e passíveis de discussão quanto aos conteúdos; acesso às fontes documentais das humanidades; uso social do conhecimento.

A partir das dez premissas apresentadas e das características das HDs, é possível refletir teoricamente sobre as possíveis relações que se apresentam no Quadro 1:

Quadro 1 - Relações entre OC e HDs

1ª Premissa: Conhecimento como produto, necessidade e dínamo social **HDs:** - uso social do conhecimento; - acesso a grande quantidade de dados; - acesso às fontes documentais das humanidades; **Possível Relação:** O conhecimento gerado na dinâmica social é ao mesmo tempo o registro das ações humanas e provisório, no sentido de pertencer a um determinado contexto. Infere-se que, junto às características das HDs, pode ser preservado e disponibilizado no sentido de atender às necessidades sociais e gerar novos conhecimentos.
2ª Premissa: Conhecimento se dá a partir da informação e quando socializado, volta a ser informação;

HDs: - uso social do conhecimento. **Possível Relação:** O uso social do conhecimento se relaciona com a premissa na direção de refletir que cada informação tem sentidos diferentes que serão compreendidas de acordo com o contexto. Para que a informação se transforme em conhecimento, é necessária a atenção às particularidades contextuais que fazem parte do ambiente e do universo dos usuários.
3ª Premissa: A estrutura e a comunicação do conhecimento formam um sistema aberto; **HDs:** - acesso a grande quantidade de dados; - caráter colaborativo; - acesso às fontes documentais das humanidades; - uso social do conhecimento. **Possível Relação:** O conhecimento pode ser observado a partir de perspectivas de como se forma, como se organiza, como é transmitido e utilizado. Também deve-se ter atenção ao fato de que nenhum conhecimento é estático. A partir das relações dispostas, infere-se que a OC pode contribuir com critérios, diretrizes e modelos para melhor acesso e uso desse conhecimento a partir das discussões das HDs. Por exemplo, pode ser estudado como se adquire, como se organiza, como se transmite, como é usado, quantos tipos de conhecimento existem, como se modifica ou se torna obsoleto, entre outros. Assim, acredita-se que a Organização do Conhecimento pode fornecer critérios, diretrizes e procedimentos para tais estudos, visto que o conhecimento muda conforme as necessidades sociais.
4ª Premissa: O conhecimento deve ser organizado para seu melhor aproveitamento individual e social; **HDs:** - acesso a grande quantidade de dados; - caráter colaborativo; - acesso às fontes documentais das humanidades; - uso social do conhecimento. **Possível Relação:** O conhecimento é diverso e exige um amplo repertório de possibilidades de organização para que possa ser acessado. Infere-se que a partir dos preceitos do Sistema de Conceitos relacionado às caraterísticas das HDs, deve-se trabalhar em prol da possibilidade de compreensão e disponibilização desse conhecimento em diferentes níveis de comunicação.
5ª Premissa: Existem várias formas possíveis de se organizar o conhecimento; **HDs:** - acesso a grande quantidade de dados; - caráter colaborativo; - acesso às fontes documentais das humanidades; - uso social do conhecimento. **Possível Relação:** Considerando a multiplicidade de organização do conhecimento e as várias maneiras ou disciplinas que podem acessá-lo, infere-se que a relação aqui se dá através das prováveis possibilidades e abordagens mais parciais ou gerais de organização, para acesso de diferentes usuários pertencentes a diferentes áreas do conhecimento.
6ª Premissa: Toda organização do conhecimento é provisória, artificial e determinista; **HDs:** - acesso a grande quantidade de dados; - acesso às fontes documentais das humanidades; - uso social do conhecimento. **Possível Relação:** A partir do determinismo ocidental, herança da antiguidade, e das técnicas desenvolvidas ao longo do tempo, a OC se mostra artificial e temporária, pois é uma construção abstrata e, ligada à ciência, está em desenvolvimento constante. Infere-se que o olhar deve se voltar às HDs no sentido de perceber estas determinações e elaborar Sistemas de Organização do Conhecimento alinhados à características mais humanísticas, críticas e éticas.
7ª Premissa: O conhecimento é sempre registrado em documentos, se monstra como conjunto de dados organizados e permite usos indiscriminados; **HDs:** - ferramentas de aprendizagem; - acesso a grande quantidade de dados; - caráter colaborativo; - acesso às fontes documentais das humanidades; - uso social do conhecimento. **Possível Relação:** Barite (2001, p. 51), escreve que "Los fondos documentales que administramos son también realidades objetivas [...]". Infere-se que, na relação com as HDs, deve-se considerar objetivo de preservar, organizar e disponibilizar os recursos informacionais de modo indiscriminado para a socialização do conhecimento.
8ª Premissa: O conhecimento se expressa em conceitos e se organiza a partir de sistemas de conceitos; **HDs:** - acesso às fontes documentais das humanidades; - uso social do conhecimento. **Possível Relação:** Os conceitos possuem camadas de significação, sendo construídos a partir da dinâmica social e carregando relações contextuais. Infere-se que, a partir do Sistema de Conceitos e das premissas da OC, a relação com as HDs deve também se atentar à estruturação de áreas e subáreas que representem o conhecimento que será disponibilizado, refletindo as tensões e contradições conceituais dispostas.
9ª Premissa: Os sistemas de conceitos são organizados para fins científicos, sistemáticos ou de documentação; **HDs:** - ferramentas de aprendizagem; - acesso a grande quantidade de dados; - caráter colaborativo; - acesso às fontes documentais das humanidades; - uso social do conhecimento. **Possível Relação:** Infere-se que os Sistemas de Organização do Conhecimento têm limitações referentes a organização e linguagem que, diante das HDs, podem ser pensadas para fins específicos, considerando-se as relatividades temporais e espaciais, o determinismo temático e os esquemas e discursos direcionados aos usuários.
10ª Premissa: As leis que regem a organização de sistemas de conceitos são uniformes e previsíveis e se aplicam igualmente a qualquer área disciplinar. **HDs:** - ferramentas de aprendizagem; - acesso a grande quantidade de dados; - caráter colaborativo; - acesso às fontes documentais das humanidades; - uso social do conhecimento. **Possível Relação:** Barite (2001), explica que essa lei tem uma importância fundamental para a organização do conhecimento. Infere-se aqui que, a partir das premissas e características da HDs, é importante pensar nos processos de tratamento da informação e do conhecimento para que sejam disponibilizados. Relações hierárquicas, possiblidades de associação de conceitos, observação de sinonímias e a análise do conceito desde o estado morfológico ao semântico são pontos estruturais nessa relação.

Fonte: Dados da pesquisa (2022).

A partir dos resultados apresentados, os destaques podem ser verificados, por exemplo, quando OC apresenta o relacionamento conceitual em relação a um determinado domínio, enquanto as HDs prezam por terem apelo ao acesso a grande quantidade de dados, serem colaborativas e oferecerem possibilidade de acesso às fontes documentais das humanidades. Todos esses elementos se relacionam nas premissas no

sentido de envolver o rigor metodológico de construção de Sistemas de Organização do Conhecimento, onde são preservados os sentidos dos domínios e de suas comunidades discursivas com o fim no usuário.

Um elemento comum é o uso social do conhecimento. Considerando que todas as premissas delimitam um relacionamento com os usuários e uma garantia de que os conteúdos e documentos tratados serão coerentes com o contexto apresentado pelas comunidades discursivas, a OC, como um conjunto de processos que servirão para a organização e representação das informações, e as HDs com o foco em aprendizagem, assim como o conhecimento e competências dos usuários em relação a novas metodologias e reflexões quanto a área de humanidades, podem ser percebidas em seu princípio, como ferramentas específicas direcionadas a socialização do conhecimento, visto que são colaborativas. Os processos da OC refletidos nas premissas propostas, fornecem as bases classificatórias, taxonômicas e de relacionamento conceitual, enquanto as HDs apresentam os documentos digitalizados, as fontes documentais e o conteúdo propriamente dito, que necessita de tratamento para poder ser disponibilizado.

Nesse sentido, as interlocuções a partir premissas apresentadas, enfatizam a importância da observação e estudo acerca de elementos que influenciarão diretamente na organização e na recuperação da informação, considerando as especificidades e as diferenças sociais e contextuais, que são o espaço onde a informação é efetivamente construída e compartilhada, nesse caso, na instância das humanidades.

A combinação dos interesses das Humanidades Digitais quanto as fontes necessárias para se pensar um domínio específico e dar acesso ao conhecimento, com os objetivos dos processos da OC, apresentados pelas premissas de Barite (2001), demonstra que há um caminho para direcionar de forma coerente e crítica as necessidades de diferentes usuários e assim, conduzir a maior acessibilidade e maneiras de gerir, disseminar e criar novos conhecimentos.

5 Considerações Finais

A articulação de conhecimentos no universo digital e das Ciências Sociais e Humanas é elemento de conciliação, quando observada a centralidade na digitalização e disponibilização de fontes, mas também a procura da sistematização e análise dessas fontes, para a recuperação e compreensão de forma adequada aos usuários e seus contextos.

Nesse sentido, os processos da Organização do Conhecimento permitem o controle da linguagem utilizada nos espaços de informação, como os ambientes digitais, onde os objetos passam a um estado múltiplo de fixação, ou seja, os registros informacionais podem estar em mais de um lugar ao mesmo tempo, fazendo com que as informações não fiquem armazenadas de forma contígua. Assim, a reflexão teórica sobre as premissas apresentadas por Barité (2001), combinadas com as características das HDs, têm o potencial de elaboração de um controle nas relações conceituais, além de materializar alguns objetivos das HDs, como o de disponibilizar e disseminar o conhecimento por meio digital, fornecer bases integradas às Ciências Humanas e Sociais no sentido de metodologias e teorias e potencializar as atividades digitais diante das perspectivas da humanização visando os usuários, seus relacionamentos com os conceitos e contextualização dos documentos. Neste sentido, as premissas podem fornecer um elo para o início das discussões sobre as interlocuções entre OC e as HDs.

As relações dispostas nas premissas da OC encontram uma interlocução com as HDs em diversos momentos e abarcam projetos e perspectivas interdisciplinares, que vão ao encontro da necessidade de reflexão mais profunda em torno dos aspectos, como a produção de registros digitais para a recuperação e disseminação do conhecimento, indexação, classificação e representação de forma coerente do conhecimento no âmbito das Humanidades de forma geral.

Referências

BARITÉ, M. Organizacion del conocimiento: un nuevo marco teorico-conceptual en bibliotecologia y documentacion. In:

CARRARA, K. (Org.). **Educação, universidade e pesquisa**: textos completos do III simpósio em filosofia e ciência: paradigmas do conhecimento no final do milênio. Marília: Unesp-Marília-Publicacoes; São Paulo: FAPESP, 2001. p. 35-60.

ALBUQUERQUE, A. C. de; HERNANDEZ QUINTANA, A. R. Sistemas de Organização do Conhecimento e Humanidades Digitais: possíveis interlocuções a partir da abordagem da análise do domínio. In: SILVA, C. G. da; REVEZ, J.; CORUJO, L. (Coord.). Organização do Conhecimento no Horizonte 2030: Desenvolvimento Sustentável e Saúde, Atas do V Congresso ISKO Espanha-Portugal. Lisboa: Centro de Estudos Clássicos, Faculdade de Letras, Universidade de Lisboa, 2021, p. 727-737.

ALMEIDA, M. A.; DAMIAN, I. P. M. Humanidades digitais: um campo praxiológico para mediações e políticas culturais? In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16, 2015, João Pessoa. Anais... João Pessoa: Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação, 2015. Disponível em: http://www.ufpb.br/evento/lti/ocs/index.php/enancib2015/enancib2015/paper/view/2999/1046. Acesso em: 09 abr. 2022.

ANDRADE, L. M. de; DAL'EVEDOVE, Paula Regina. Aproximações entre organização do conhecimento e humanidades digitais, In: SILVA, C. G. da; REVEZ, J.; CORUJO, L. (Coord.). Organização do Conhecimento no Horizonte 2030: Desenvolvimento Sustentável e Saúde, Atas do V Congresso ISKO Espanha-Portugal. Lisboa: Centro de Estudos Clássicos, Faculdade de Letras, Universidade de Lisboa, 2021, p. 739-749.

CASTRO, R. M., PIMENTA, R. M. (2018). Novas práticas informacionais frente às humanidades digitais: a construção de acervos digitais como suporte para as digital humanities. **Informação & Informação**, v. 23, n. 3, p. 523-543, 2018. Disponível em: http://dx.doi.org/10.5433/1981-8920.2018v23n3p523. Acesso em: 20 abr. 2022

GALINA RUSSELL, I. (2011). ?Qué són las humanidades digitales?. **Revista Digital Universitaria**, v. 12, b. 7, 2011. Disponível em: http://eprints.rclis.org/19368/1/037-043-Rz-Yunta-Humanidades-digitales.pdf. Acesso em: 09 abr. 2022.

HJØRLAND, B.Concept theory. **Journal of the American Society for Information Science and Technology**, v. 60, n. 8, p. 1519-1536, 2009. Disponível em: https://doi.org/10.1002/asi.21082. Acesso em: 12 abr. 2022.

HOCKEY, S. The history of humanities computing. In: SCHREIBMAN, S.; SIEMENS, R.; UNSWORTH, J. (Eds). **A companion to digital humanities**. Malden, MA: Blackwell Publishing. p. 3-19.

KOH, A. **A letter to the humanities: dh will not save you.** 2015. Disponível em: http://www.digitalpedagogylab.com/hybridped/a-letter-to-the-humanities-dh-will-not-saveyou/. Acesso em; 09 abr. 2022.

MANIFESTO DE HUMANIDADES DIGITAIS, 2011. Disponível em: https://humanidadesdigitais.org. Acesso em: 12 abr. 2022.

MOURA, M. Interoperabilidade Semântica e Ontologia Semiótica: a construção e o compartilhamento de conceitos científicos em ambientes colaborativos online. **Informação & Informação,** *v. 16, n.* 2, p. 165-179, maio/ago. 2011. Disponível em: doi: http://dx.doi.org/10.5433/1981-8920.2011v16n2p165. Acesso em: 10 abr. 2022.

MOURA, M. Ciência da Informação e humanidades digitais: mediações, agência e compartilhamento de saberes. *Perspectivas em Ciência da Informação*, v. *24, n.* esp., p. 57-69, 2019. Disponível em: https://doi.org/10.1590/1981-5344/3893. Acesso em: 10 abr. 2022.

PALETTA, F. C. Ciência da Informação e humanidades digitais: uma reflexão. 2018. In XIX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO. p. 147-162. Londrina, PR.

PIMENTA, R. M. As rugosidades do Ciberespaço: um contributo teórico aos estudos dos web espaços informacionais. **Informação & Sociedade: Estudos**, v. *26, n.* 2, p. 77-90, maio/ago., 2016. Disponível em: https://periodicos.ufpb.br/index.php/ies/article/view/28116. Acesso em: 09 abr. 2022.

PIMENTA, R. M. Por que Humanidades Digitais na Ciência da Informação? Perspectivas pregressas e futuras de uma prática transdisciplinar comum. **Informação & Sociedade: Estudo***s*, v. *30, n.* 2, maio/ago., 2020. Disponível em: https://doi.org/10.22478/ufpb.1809-4783.2020v30n2.52122. Acesso em: 09 abr. 2022.

TANG, M. C.; CHENG, Y. J.; CHEN, K. H. Um estudo longitudinal de coesão intelectual em humanidades digitais usando análises bibliométricas. **Scientometrics,** v. 113, p. 985–1008. Disponível em: https://doi.org/10.1007/s11192-017-2496-6. Acesso em: 09 abr. 2022

A WEB SEMÂNTICA APLICADA NA RECUPERAÇÃO DE INFORMAÇÃO: UM ESTUDO DE CASO NO CONTEXTO ESTATÍSTICO DE USO DE LIVROS DIGITAIS POR ALUNOS DE GRADUAÇÃO

THE SEMANTIC WEB APPLIED IN INFORMATION RETRIEVAL: A CASE STUDY IN THE STATISTICAL CONTEXT OF THE USE OF DIGITAL BOOKS BY UNDERGRADUATE STUDENTS

Stella Schwanz Dias de Assis[1], Alessandra Monteiro Pattuzzo Caetano[2], Henrique Monteiro Cristovão[3]

(1) Universidade Federal do Espírito Santo - UFES, Vitória/ES, e-mail:stella.assis@edu.ufes.br

(2) Universidade Federal do Espírito Santo - UFES, Vitória/ES, e-mail: apattuzzo@gmail.com

(3) Universidade Federal do Espírito Santo - UFES, Vitória/ES, e-mail: henrique.cristovao@ufes.br

Resumo

Considerando o contexto dos dados estatísticos de uso de livros digitais da biblioteca virtual de uma instituição de ensino superior por seus usuários alunos de graduação, e considerando também o contexto da recuperação de informação na Web semântica em dados ligados RDF, a presente pesquisa tem como objetivo mostrar o processo de mapeamento dos dados existentes, em uma pequena parte da base dessa biblioteca, para dados ligados na Web semântica, dando ênfase ao estabelecimento de interoperabilidade nas camadas sintática, estrutural e semântica. Como prova de conceito, foram executadas consultas sobre a base de dados criada para recuperar informações. Usou-se abordagem qualitativa, natureza aplicada e procedimentos de estudo de caso. Seguindo um *workflow* composto de 10 etapas, desde o conhecimento dos dados e contexto, passando pela elaboração de uma modelagem conceitual simplificada, limpeza, preparação e reconciliação de dados, implementação de uma ontologia operacional, mapeamento RDF, exibição de *knowledge graphs*, até a criação de uma base de dados ligados e a execução de consultas escritas em SPARQL, o percurso da pesquisa conseguiu cumprir o objetivo proposto. Como o desenvolvimento foi realizado em extrato pequeno da base de dados, há necessidade de ampliar a pesquisa, inclusive com a elaboração de uma ontologia de domínio completa para o cenário, bem como a implementação de interface para uso da linguagem de consulta SPARQL, uma vez que ela não é apropriada para o usuário final.

Palavras-chave: Recuperação da informação; Web semântica; Interoperabilidade; Acesso à informação; Livros digitais.

Abstract

Considering the context of statistical data on the use of digital books from the virtual library of a higher education institution by its undergraduate students, and also considering the context of information retrieval on the semantic Web in data linked to RDF, this research aims to show the process of mapping existing data, in a small part of the base of this library, to linked data in the semantic Web, emphasizing the establishment of interoperability in the syntactic, structural and semantic layers. As a proof of concept, queries were performed on the database created to retrieve information. A qualitative approach was used with an applied purpose and case study procedures. Following a workflow composed of 10 steps, from knowing the data and context, passing through the elaboration of a simplified conceptual model, cleaning, preparation and reconciliation of data, implementation of an operational ontology, RDF mapping, display of knowledge graphs, until the creation of a linked database and the execution of queries written in SPARQL, the research course was able to fulfill the proposed objective. As the development was carried out in a small extract of the database, there is a need to expand the research, including the elaboration of a complete domain ontology for the scenario, as well as the implementation of the interface for using the SPARQL query language, since that it is not appropriate for the end user.

Keywords: Information retrieval; Semantic Web; Interoperability; Access to information; Digital books.

1 Introdução

A recuperação de informação (RI) trata da representação, armazenamento, organização e acesso a itens de informação, como documentos, páginas da Web, catálogos online, registros estruturados e semiestruturados, objetos multimídia. A representação e organização dos itens de informação devem ser tais que proporcionem aos usuários fácil acesso às informações de

seu interesse (BAEZA-YATES; RIBEIRO-NETO, 2011).

Materializando a Web semântica, os dados ligados surgem como um conjunto de boas práticas para publicar e conectar conjuntos de dados estruturados na Web, com intuito de criar uma Web de dados ligados (BIZER; HEATH; BERNERS-LEE, 2009).

Nesse contexto de demandas da RI a Web semântica surgiu com o intuito de melhorar a organização da Web e, consequentemente, tornar os seus dados mais facilmente localizáveis inserindo semântica nesses dados com o objetivo que eles sejam melhores compreendidos pelas máquinas (BERNERS-LEE; HENDLER; LASSILA, 2001). Além disso, considera-se que o modelo de dados ligados é unificado e projetados para o compartilhamento global (HEATH; BIZER, 2011) por meio de um planejamento adequado de interoperabilidade das camadas sintática, estrutural e semântica (ZENG, 2019). Interoperabilidade é definido pela National Information Standard Organization (NISO) como "[...] a capacidade de múltiplos sistemas com diferentes hardwares e softwares plataformas, estruturas de dados e interfaces para troca de dados com o mínimo perda de conteúdo e funcionalidade" (NISO, 2004, p. 2, tradução nossa).

A oferta de bibliotecas virtuais vem aumentando no país, e as instituições de ensino superior precisam inovar o processo de acesso e disponibilização da informação. Este é um cenário que não tem volta, e os profissionais bibliotecários, além de disponibilizar, precisam apresentar e acompanhar o uso efetivo dos livros digitais para tomadas de decisão em planejamentos administrativos, financeiros e pedagógicos em suas unidades de informação. Além disso, em um contexto de distribuição de informações na Web, a tarefa de seleção e consulta destas se tornou algo impossível de ser realizado por pessoas, o que torna evidente a necessidade da utilização de aplicações com essa destinação. Nesse sentido, não são estipuladas regras para publicação dos dados, mas apresentadas boas práticas que possibilitam a criação de modelos de dados capazes de se comunicar que sejam interoperáveis. A Web semântica se posiciona como uma extensão da Web (SERRA, 2019).

Em relação ao desenvolvimento de coleções, as decisões de cancelar ou permanecer com o acervo virtual, ocorre durante seu tempo de vigência por meio de controle estatístico de uso e se o acervo disponibilizado é viável para a composição das referências bibliográficas dos planos de ensino das disciplinas dos cursos ofertados pela Instituição de Ensino Superior (IES).

Observa-se uma emergencial necessidade de geração de indicadores que permitam aos profissionais, que atuam em bibliotecas, a elaboração de relatórios e estatísticas de uso dos livros virtuais para a criação de um cenário econômico sustentável pela IES e para dar suporte pedagógico na elaboração dos documentos educacionais essenciais para os funcionamentos dos cursos ofertados pela IES.

Dessa forma, o presente trabalho se propõe a responder ao seguinte problema de pesquisa: como se utilizar dos recursos da Web semântica para organizar os dados disponíveis nas bases de uma biblioteca a fim de fornecer relatórios e estatísticas que auxiliam a gestão da IES ao qual ela pertence?

2 Objetivos

Considerando o cenário dos dados e variáveis existentes na base de dados da biblioteca do estudo de caso, a presente pesquisa pretende mostrar, para um pequeno extrato dessa base, o processo de mapeamento dos dados existentes para dados ligados na Web semântica dando ênfase ao estabelecimento de interoperabilidade nas camadas sintática, estrutural e semântica, e uma prova de conceito para a recuperação de informações a partir de uma linguagem de consulta específica de dados ligados.

3 Procedimentos Metodológicos

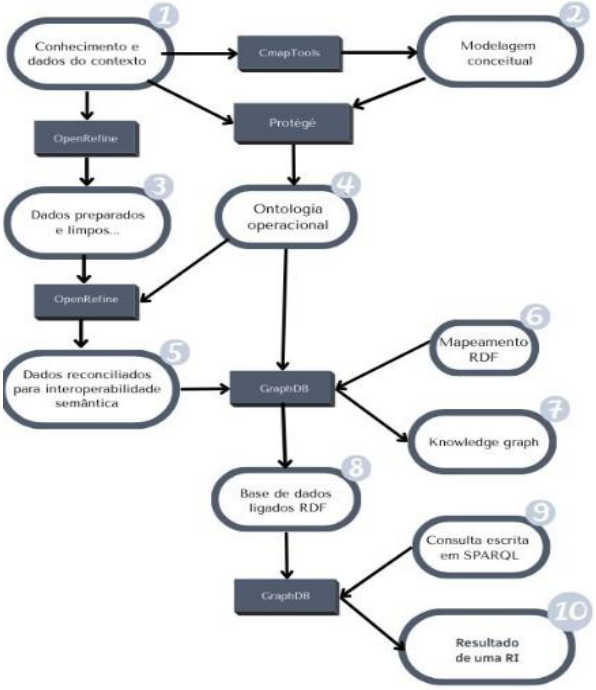
O presente estudo usa abordagem qualitativa, natureza aplicada, sendo utilizado também procedimentos de estudo de caso, que segundo Yin (2001) proporciona um processo investigativo onde são preservadas as características holísticas e significativas dos eventos da vida real.

Com o objetivo de testar a proposta desenvolvida utilizou-se uma prova de conceito pela escrita de consultas de forma a obter um resultado composto de dados de outras bases interligadas por meio de metadados interoperáveis.

O contexto delimitado para a pesquisa está nos relatórios de acesso da biblioteca digital da FAESA - Centro Universitário Espírito Santense, sendo que os dados foram gerados por relatórios da Vital Source[1], provedora do acervo digital, referentes aos acessos do ano de 2021 no primeiro semestre (janeiro a junho).

O *workflow* do desenvolvimento realizado, mostrado na Figura 1, é composto de 10 caixas numeradas que correspondem a etapas, ou elementos informacionais, que alimentam ou que são gerados pelas ferramentas computacionais indicadas pelas caixas escuras, cujas direções de interação estão sinalizadas por setas. Essas 10 etapas do *workflow* são explicados nos próximos parágrafos.

Figura 1 - Workflow do desenvolvimento



Fonte: autoria própria

Etapa 1 - Conhecimento e dados do contexto. Correspondeu ao levantamento dos dados existentes e conhecimento do funcionamento do contexto.

Etapa 2 - Modelagem conceitual. Considerando a redução do cenário dos dados e variáveis existentes na base, foi realizada uma modelagem conceitual simplificada indicando-se as principais entidades do contexto e seus relacionamentos, com o uso da ferramenta CmapTools[2].

Etapa 3 - Dados preparados e limpos. Os dados foram analisados, limpos e preparados com apoio do software OpenRefine[3].

Etapa 4 - Ontologia operacional. A partir da modelagem conceitual e com apoio do software Protégé[4], foi gerada uma ontologia operacional, que segundo Falbo (2014), em sua metodologia para construção de ontologias, denominada de SABiO, corresponde a implementação da ontologia de domínio diretamente em uma linguagem operacional. No caso da presente pesquisa, no contexto dos dados ligados em RDF (Resource Description Framework), foi escolhida a linguagem RDF Turtle[5]. Ainda nessa etapa foram escolhidos elementos interoperáveis da camada sintática e estrutural para incorporação na linguagem. Nessa etapa também buscou-se atender a camada semântica da interoperabilidade pela equivalência de classes.

Etapa 5 - Dados reconciliados para interoperabilidade semântica. Utilizando-se do serviço de reconciliação de dados do OpenRefine, foi executada sobre algumas variáveis escolhidas onde se achou dados abertos correspondentes para acesso. Essa etapa cumpriu, ainda que de forma pequena, a agregação de dados para compor a

[1] VitalSource está disponível em: https://www.vitalsource.com/.

[2] CmapTools é um software para representação de conhecimento. Disponível em: https://cmap.ihmc.us/cmaptools/.

[3] OpenRefine é um software para limpeza, preparação e reconciliação de dados. Disponível em: https://openrefine.org/.

[4] Protégé é um editor de ontologias de código aberto e gratuito. Disponível em: https://protege.stanford.edu/.

[5] RDF Turtle é uma linguagem de marcação para representação de dados ligados RDF. Disponível em: https://www.w3.org/TR/turtle/.

interoperabilidade semântica por via da equivalência de indivíduos.

Etapa 6 - Mapeamento RDF. Nessa etapa usou-se o software GraphDB[6] para inserir um mapeamento RDF com base na ontologia operacional gerada na etapa 4, a fim de gerar um repositório de dados ligados RDF.

Etapa 7 - Knowledge graph. Esse grafo de triplas foi gerado pelo GraphDB, usando dados reconciliados e a ontologia operacional. O intuito é apenas para conferência visual de parte da base, uma vez que é possível visualizar com facilidade as entidades e algumas instâncias com os seus respectivos relacionamentos.

Etapa 8 - Base de dados ligados RDF. Gerada pelo software GraphDB, que se utiliza do mapeamento RDF da etapa 6.

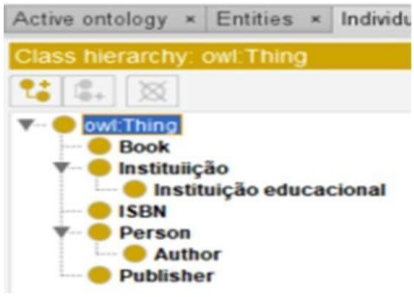
Etapa 9 - Consulta escrita em SPARQL[7]. Como prova de conceito, foram realizadas algumas consultas na base de dados ligados.

Etapa 10 - Resultado de uma RI. É o conjunto de triplas RDF advindas da consulta SPARQL realidade na etapa 9.

4 Resultados

Para a modelagem conceitual (etapa 2) selecionou-se as principais entidades do cenário: 'Book', 'Instituição educacional' (como sub categoria de 'Instituição'), 'ISBN', 'Author' (como subcategoria de 'Person'), e 'Publisher'. Em seguida, elas foram implementadas no Protégé, conforme mostra a hierarquia de classes da Figura 2, e incorporadas propriedades necessárias para criar as conexões entre os elementos bem como aspectos de interoperabilidade da camada sintática, a partir de vocabulários amplamente conhecidos como xsd[8], rdf[9], rdfs[10], skos[11], owl[12] entre outros. Também foram incorporados elementos para a camada estrutural de interoperabilidade tais como Dublin Core[13], Friend of Friend (FOAF)[14], Creative Commons(CC)[15], Bibliographic Framework Initiative (Bibframe)[16], Schema.org[17], Data Catalog Vocabulary (DCT)[18] entre outros.

Figura 2 – Hierarquia de classes da ontologia operacional



Fonte: autoria própria, tela capturada do software Protégé

O Quadro 1 mostra algumas equivalências encontradas para adição de interoperabilidade de camada semântica, correspondente à etapa 4 dos procedimentos metodológicos. A estratégia utilizada para a conexão dos dados foi pautada na inserção das propriedades nas anotações dos títulos dos livros, centralizados como principais indivíduos da modelagem.

A fim de agregar uma segunda camada de interoperabilidade semântica, não apenas conectada às classes, mas à cada um dos indivíduos, foi realizada uma reconciliação dos dados, possibilitando uma conexão com diferentes bases, onde foi possível encontrar mais informações sobre elementos

[6] GraphDB é um banco de dados orientado a grafo compatível com RDF e SPARQL. Disponível em: https://graphdb.ontotext.com/.

[7] SPARQL é uma uma linguagem de RI estruturada e padronizada para realizar consultas em grafos RDF.

[8] Prefixo e namespace: @prefix xsd: <http://www.w3.org/2001/XMLSchema#>.

[9] Prefixo e namespace: @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.

[10] Prefixo e namespace: @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.

[11] Prefixo e namespace: @prefix skos: <http://www.w3.org/2004/02/skos/core#>.

[12] Prefixo e namespace: @prefix owl: <http://www.w3.org/2002/07/owl#>.

[13] Prefixo e namespace: @prefix dcterms: <http://purl.org/dc/terms/>.

[14] Prefixo e namespace: @prefix foaf: <http://xmlns.com/foaf/0.1/>.

[15] Prefixo e namespace: @prefix cc: <http://creativecommons.org/ns#>.

[16] Prefixo e namespace: @prefix bibframe: <http://id.loc.gov/ontologies/bibframe/>.

[17] Prefixo e namespace: @prefix schema: <http://schema.org/>.

[18] Prefixo e namespace: @prefix dcat: <http://www.w3.org/ns/dcat#>.

reconciliados, desde mais informações sobre os livros através do ISBN, até a busca pelos autores e editoras.

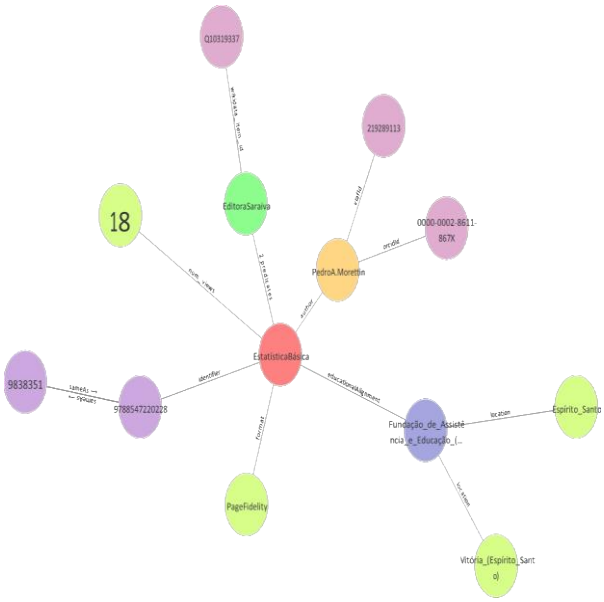
Quadro 1 - Conexões semânticas entre classes

Classe	equivalentClass
Author	https://www.wikidata.org/wi ki/Q482980
Book	https://schema.org/Book
ISBN	https://id.loc.gov/ontologies /bibframe.html#c_Isbn
Instituição Educacional	https://schema.org/Educati onalOrganization
Person	https://schema.org/Person %20 http://xmlns.com/foaf/spec/ #term_Person%20
Publisher	http://mappings.dbpedia.or g/server/ontology/pages/O ntologyClass%3APublisher

Fonte: autoria própria

A Figura 3 mostra um *knowledge graph* para conexões de um livro arbitrário, nesse caso o livro intitulado 'Estatística Básica' do autor 'Pedro A. Morettin'. Esse grafo de triplas RDF corresponde a um possível resultado da etapa 7, e utiliza-se de informações advindas da ontologia operacional da etapa 4, e de dados reconciliados da etapa 5.

FIGURA 3 - Knowledge graph para conexões de um livro



Fonte: autoria própria, gerado pelo software GraphDB

Nesse grafo da Figura 3 é possível ver relacionamentos que usam dados reconciliados e, como por exemplo, o nó do indivíduo 'Fundação de Assistência e Educação' conectado por meio da propriedade 'location' com o indivíduo 'Vitória (Espírito Santo)' advindo de uma reconciliação realizada com o software OpenRefine. Outro exemplo é o autor 'Pedro A.Morettin com o relacionamento do seu Orcid, cujo dado também veio de uma reconciliação. Há ainda conexões que estabelecem interoperabilidade semântica entre indivíduos, como é o caso do identificador do livro conectado por meio da propriedade 'sameAs' com o identificador do mesmo livro em outra base de dados.

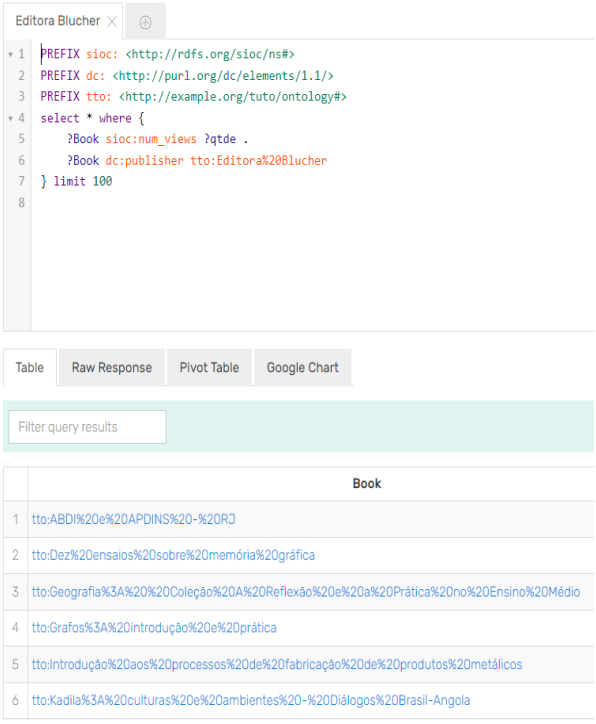
Esse tipo de representação intermediária, por meio de *knowledges graphs*, de pequenos extratos da base de dados ligados, facilita a compreensão e inspeção visual dos relacionamentos entre os dados. Também é possível refletir sobre possibilidades de ampliação de enriquecimento da base de dados ligados com elementos interoperáveis e, assim, possibilitar posteriormente uma RI que agrega elementos de outras bases seja por meio de reconciliação ou interoperabilidade semântica de indivíduos.

A prova de conceito foi obtida realizando-se algumas consultas, escritas em SPARQL, sobre a base de dados ligados no repositório criado no software GraphDB. Uma delas pode ser verificada na Figura 4 inclusive com as triplas obtidas pelo resultado da consulta. Nessa consulta da Figura 4, buscou-se os dados estatísticos da quantidade de empréstimos associados a um determinado livro.

Não obstante, é importante ressaltar também algumas dificuldades encontradas nesse processo. Como foi descrito nos procedimentos metodológicos, foi necessária a realização de uma série de filtragens para obter uma base interoperável. Existe uma carência na alimentação nas bases que possibilitam essa reconciliação, até mesmo na busca pelas editoras observou-se uma quantidade pequena, além da impossibilidade de uma reconciliação automática de alguns elementos devido a inconsistências nas informações,

normalmente por falta de atualização. Outro problema encontrado, que vale destacar, é a falta de acesso à APIs que possibilitam relacionamentos externos, como o exemplo da API fornecido pela OCLC[19] para acessar dados do WorldCat.

Figura 4 - Resultado de uma consulta SPARQL



Fonte: autoria própria, gerado pelo software GraphDB

Resgatando a necessidade informacional que desencadeou essa busca, que seria recuperar informações relevantes, por meio dos dados que levantem possibilidades de estratégias para alcançar um maior número de acessos à base, é possível propor diversas possibilidades com base na exploração e descobertas realizadas.

Uma primeira possibilidade seria reproduzir a interoperabilidade introduzida na base de dados na própria plataforma da biblioteca digital, levando esse recurso ao usuário final. Outra proposta que pode ser agregada é a consolidação de uma base de dados com a bibliográfica básica e complementar de todos os cursos na faculdade, delimitando, além dos cursos, áreas do conhecimento relacionadas e o período em que a obra é exigida. Com uma base que viabilize o acesso à essas informações, é possível enriquecer a base resultante da ontologia, agregando informações que possam contribuir para a construção de algoritmos de recomendação aplicáveis no acervo digital, relacionando os materiais essenciais para os alunos, que podem ser notificados no início de cada período sobre as obras disponíveis, além de possibilitar recomendações baseadas nesses materiais.

4 Considerações Finais

A pesquisa possibilitou a verificação positiva da utilização de elementos da Web semântica para inclusão de dados dinâmicos nos catálogos das bibliotecas digitais. Observou-se que as soluções de interoperabilidade evoluíram ao longo do tempo, possibilitando a heterogeneidade de acervos e a preservação da semântica dos conteúdos tornados interoperáveis. Observou-se também que é possível a inclusão de vínculo entre dados de um mesmo catálogo, estabelecendo relações entre registros presentes no acervo e destes com bases de dados externos e de outras ferramentas da Web semântica em bibliotecas, como o Knowledge Graph.

Foram sugeridas propostas de outras possibilidades de enriquecimento na relação interoperável dos dados, evidenciando a necessidade de realização de mais pesquisas sobre a relação entre as bibliotecas digitais e a conexão interoperável de dados da Web.

Quanto ao objetivo proposto, a pesquisa conseguiu mostrar, para um pequeno extrato da base escolhida para o estudo de caso, o processo de mapeamento dos dados existentes para dados ligados na Web semântica dando ênfase ao estabelecimento de interoperabilidade nas camadas sintática, estrutural e semântica. Foi também realizada uma prova de conceito para a recuperação de informações a partir da linguagem SPARQL.

[19] OCLC (Online Computer Library Center, Inc.) é uma organização sem fins lucrativos considerada a maior cooperativa de bibliotecas, museus e arquivos do mundo. Disponível em: https://www.oclc.org/.

Além disso, é possível inferir que, havendo mais recursos e tempo para continuidade da pesquisa, e utilizando-se o *workflow* apresentado, seria indicado contemplar a base de dados completa bem como elaborar uma ontologia de domínio que considere todo o cenário. Dessa forma, seria possível fazer um acompanhamento estatístico de uso dos livros virtuais disponíveis em uma biblioteca virtual, de forma personalizada para cada objetivo estratégico estabelecido pelo bibliotecário gestor. Seria também necessário implementar uma interface para uso da linguagem de consulta SPARQL uma vez que ela não é apropriada para o usuário final.

Referências

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Recuperação de informação**: conceitos e tecnologia das máquinas de busca. 2. ed. Porto Alegre: Bookman, 2013.

BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim. Linked data - the story so far. **International Journal on Semantic Web and Information Systems** (IJSWIS), v. 5, n. 3, p. 1–22, 2009. Disponível em: https://www.igi-global.com/article/linked-data-story-far/37496. Acesso em: 1 mar. 2021.

FALBO, Ricardo de Almeida. SABiO: Systematic approach for building ontologies. Em: 2014, Rio de Janeiro, RJ. **Anais** [...]. . Em: 1st Joint Workshop ONTO.COM/ODISE on Ontologies in Conceptual Modeling and Information Systems Engineering co-located with 8th International Conference on Formal Ontology in Information Systems. Rio de Janeiro, RJ: CEUR Workshop Proceedings, 2014. p. 14. Disponível em: http://ceur-ws.org/Vol-1301/ontocomodise2014_2.pdf. Acesso em: 13 nov. 2021.

HEATH, Tom; BIZER, Christian. Linked data: evolving the web into a global data space. **Synthesis Lectures on the Semantic Web**: Theory and Technology, v. 1, n. 1, p. 1–136, 2011. Disponível em: https://www.morganclaypool.com/doi/abs/10.2200/S00334ED1V01Y201102WBE001. Acesso em: 2 mar. 2021.

NISO. **Understanding metadata**. Bethesda (EUA): National Information Standards Organization, 2004. Disponível em: http://www.niso.org/publications/press/UnderstandingMetadata.pdf . Acesso em: 13 nov. 2021.

SERRA, Liliana Giusti. A web semântica na gestão de livros digitais licenciados: uma proposta de modelo. 2019. Tese (Doutorado em Ciência da Informação), Programa de Pós-Graduação em Ciência da Informação da UNESP, São Paulo, 2019. Disponível em:https://repositorio.unesp.br/handle/11449/183526. Acesso em: 02 nov. 2021.

YIN, Robert K. **Estudo de caso**: planejamento e métodos. 3. ed. Porto Alegre, RS: Bookman, 2001

ZENG, Marcia Lei. Interoperability. **Knowledge Organization**, v. 46, n. 2, p. 122–146, 2019. Disponível em: http://www.isko.org/cyclo/interoperability. Acesso em: 13 nov. 2021.

ANÁLISE DA REGIONALIDADE DO CONJUNTO DE ARTIGOS PUBLICADOS EM EVENTOS CIENTÍFICOS

ANALYSIS OF THE REGIONALITY OF THE SET OF ARTICLES PUBLISHED IN SCIENTIFIC EVENTS

Fernanda Silva Coimbra[1], Thiago Magela Rodrigues Dias[2], Ronaldo Ferreira de Araújo[3]
(1) CEFET- MG, coimbra.sfernanda@gmail.com
(2) CEFET-MG, thiagomagela@cefetmg.br
(3) UFAL, ronaldo.araujo@ichca.ufal.br

Resumo
A maioria dos trabalhos que avaliam a produção científica em eventos científicos possui repositórios específicos como fonte de dados, muitas vezes restritos a algumas áreas do conhecimento. Embora tais trabalhos apresentam resultados interessantes, não foram encontrados estudos macro que englobam os eventos em geral, com grande número de publicações ou que abordam diversas áreas. Nesse contexto, este trabalho tem como objetivo verificar características da produção científica publicada em eventos, compreendendo a regionalidade destes eventos, utilizando métricas bibliométricas nos dados extraídos dos currículos cadastrados na Plataforma Lattes de um grupo de 360.888 doutores, ao todo foram analisados 11.416.655 trabalhos publicados em anais de eventos. Os resultados demonstram uma análise da regionalidade dos eventos no âmbito de países, estados e cidades. Possibilitando verificar que a maioria dos trabalhos são em eventos no Brasil, predominantemente na região Sudeste e Sul, também é possível compreender as cidades que tiveram mais eventos científicos.

Palavras-chave: Produção científica; eventos científicos; anais; Plataforma Lattes.

Abstract
Most works that evaluate scientific production in scientific events have specific repositories as a source of data, often restricted to some areas of knowledge. Although such works present interesting results, no macro studies were found that encompass events in general, with a large number of publications or that address several areas. In this context, this work aims to verify characteristics of scientific production published in events, including the regionality of these events, using bibliometric metrics in the data extracted from the curricula registered in the Lattes Platform of a group of 360.888 doctors, in all 11.416.655 works were analyzed published in annals of events. The results demonstrate an analysis of the regionality of the events in the scope of countries, states and cities. Making it possible to verify that most of the works are in events in Brazil, predominantly in the Southeast and South regions, it is also possible to understand the cities that had more scientific events.

Keywords: Scientific production; scientific events; annals; Lattes Platform.

1 Introdução

A produção científica é propagada de diversas formas e através da *internet* é possível ter acesso aos trabalhos científicos. Ao longo dos últimos anos vários autores têm se interessado pela análise da produção científica, aumentando assim o volume de estudos que visam entender como a ciência evolui e de qual forma acontece a colaboração científica (Dias, 2016).

As produções científicas são partes integrantes do processo de produção de conhecimento do indivíduo, no qual pode disponibilizar esse conhecimento adquirido através diversas formas (Domingues, 2014). Um exemplo de meio de propagação do conhecimento são os eventos científicos, nestes eventos são gerados documentos, popularmente conhecidos como anais. As publicações geradas nos eventos são vistas por alguns estudos, como as produções acadêmicas mais atuais (Carmona; Pereira, 2018).

Analisar como as publicações neste meio de divulgação vem sendo realizadas, se apresenta como um importante mecanismo para a compreensão da evolução dos eventos científicos em um contexto geral ou em determinadas áreas do conhecimento.

Porém, as informações relacionadas à produção científica por meio de eventos científicos estão presentes em inúmeros

repositórios de dados, dificultando assim a recuperação e análise dos dados de forma ampla. Representando desta forma um grande desafio a ser enfrentado para este tipo de análise.

Neste contexto, no Brasil, a Plataforma Lattes do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) se tornou um padrão para registro de dados curriculares da comunidade científica. A Plataforma Lattes se transformou no local onde os estudantes e pesquisadores do país podem registrar sua vida pregressa e atual, ou seja, um padrão nacional, com uma riqueza de informações, confiabilidade e abrangência (Plataforma Lattes, 2022). A Plataforma Lattes tem como premissa que os indivíduos realizem a inserção de todas suas informações curriculares na sendo que, após a inclusão, todos estes dados estão disponíveis em acesso aberto na internet, podendo ser editados no momento que o indivíduo optar. É um rico repositório, contemplando o registro da trajetória profissional, acadêmica e produção científica, no qual possibilita diversas e diferentes análises, justificando assim, a escolha desse repositório como fonte de dados para este trabalho. Porém, não é possível recuperar todos os currículos de uma única vez.

Desse modo, a Plataforma Lattes surge como uma ótima alternativa para coleta de dados do conjunto de dados sobre publicações científicas em anais de eventos.

Logo, com o conjunto de dados sobre os artigos publicados em anais de eventos registrados nos currículos da Plataforma Lattes, diversas métricas bibliométricas poderão ser aplicadas para verificar características da regionalidade da produção científica brasileira neste meio de comunicação.

Logo, o principal objetivo deste trabalho é a análise do conjunto de dados curriculares da Plataforma Lattes, no intuito de se apresentar informações inéditas sobre este meio de publicação no Brasil.

2 Procedimentos Metodológicos

A metodologia utilizada neste estudo teve como base a análise bibliométrica concomitante com metodologia quantitativa. Foi utilizado o *LattesDataXplorer* (Dias, 2016), esse arcabouço permite realizar o processo de extração e seleção dos dados curriculares da Plataforma Lattes, que envolve um conjunto de técnicas e métodos, que possibilita a coleta, seleção, tratamento e análise dos dados. Para este trabalho, foi utilizado apenas os módulos de coleta e seleção do *LattesDataXplorer*, com o intuito de extrair e selecionar os dados curriculares da Plataforma Lattes. Deste modo, o módulo Coleta, foi realizado em etapas, sendo elas: coleta de URLs (*Uniform Resource Locator*), coleta de identificadores e extração dos currículos.

O extrator *LattesDataXplorer* foi utilizado em janeiro de 2021, para realizar a coleta de todos os currículos, cerca de 7 milhões de registros. O formato dos currículos é XML (*eXtensible Markup Language*), tal formato permite delimitações e é apropriado para o processamento automático, possibilitando uma melhor manipulação dos dados.

Para obter uma análise acurada dos artigos publicados em anais de eventos, preferiu-se estabelecer um conjunto de indivíduos através do nível de formação acadêmica/titulação que possuem o nível de formação doutorado concluído. A escolha é feita mediante ao que Dias (2016) menciona em seu trabalho: os doutores são responsáveis por 74,51% dos artigos publicados em periódico e 64,67% dos artigos publicados em anais de eventos, além de possuir em geral data de atualização de seus currículos recente e notadamente são responsáveis pelo mais alto nível de formação. Com o intuito de auxiliar as análises dos dados, após a utilização do arcabouço, foi necessário criar métodos para a seleção, tratamento e visualização conforme Figura 1 (Apêndice 01).

A etapa de tratamento baseou-se em analisar o conjunto de arquivos XML. Para cada currículo foi extraído um arquivo em formato XML. O currículo XML extraído da Plataforma Lattes apresenta sua estrutura o elemento raiz, denominado como “Curriculo”; e possui cinco elementos filhos que possuem seus próprios elementos e atributos. Cada currículo é único e possui informações próprias; esses dados agregam informações sobre grandes áreas, formação acadêmica, orientações, produções, entre outras. Nesta etapa da análise é realizada a divisão das informações do currículo, posteriormente

acessar as informações de interesse e desprezar algumas informações que para este trabalho são irrelevantes. Após realizar o tratamento do XML, foi realizada a etapa de visualização em que é realizada a caracterização dos dados, permitindo a análise dos dados inseridos na Seção de Trabalhos em Eventos, seção que contém trabalhos publicados em anais de eventos, caso o indivíduo tenha informado em seu currículo.

Diversos são os desafios enfrentados nas etapas de tratamento e transformação dos dados, em especial, a falta de padrão no registro dos títulos dos eventos científicos. Para contornar tal desafio, a utilização de algumas técnicas de Processamento de Linguagem Natural foram adotadas, como por exemplo a identificação de similaridades entre *strings*, e ainda, a adoção de dicionários controlados para validação dos dados.

Através da caracterização dos dados foi possível obter indicadores gerais como: total de artigos publicados em anais de eventos, ano de publicação dos artigos, totais de indivíduos que possuem artigos publicados em anais de eventos, publicações por grande área, indivíduos sem publicações, publicações com identificadores persistentes; tais indicadores são apresentados na seção de Resultados.

3 Resultados

A caracterização inicial resultou em um conjunto de 360.888 currículos, este quantitativo representa aproximadamente 5% de currículos cadastrados na Plataforma Lattes (dados de janeiro de 2021). Dentre esses currículos foi verificado que 57.403 currículos não possuíam nenhum artigo informado na seção de trabalhos publicados em anais de eventos, correspondendo a 16% do conjunto de doutores. Após as análises dos dados foi possível compreender algumas características no que tange a regionalidade dos eventos científicos.

3.1 Regionalidade dos eventos científicos

Em todos os trabalhos inseridos na Plataforma Lattes é possível informar o "País do Evento", se tornando uma informação relevante para as análises dos eventos científicos. A produção científica brasileira também pode ser publicada em eventos internacionais. Foram encontrados 185 países vinculados aos trabalhos analisados. Devido a este volume de países, foi realizado o ranking com os 10 países que possuem maior número de trabalhos publicados em anais de eventos, conforme Figura 2 (Apêndice 02).

Pode-se observar que o Brasil é o país que possui maior número de publicações. A falta de investimento para tais publicações internacionais, que em geral, necessitam de recursos para viagens, pode ser uma hipótese. Segundo Serra, Fiates e Ferreira (2008), os investimentos das instituições de ensino superior para enviar seus docentes a eventos internacionais vêm diminuindo rapidamente. Os critérios de participação entre congressos nacionais e internacionais são de suma importância para a melhoria da produção acadêmica, devendo ser olhado como investimento e não como custo.

Através das análises realizadas, foi possível constatar que na Plataforma Lattes, possui trabalhos publicados em anais de eventos em 185 países, abrangendo praticamente todos os continentes: África, América do Norte, América do Sul, Ásia, Europa e Oceania. Mesmo sendo predominante os eventos serem no Brasil, pode-se observar, que a produção científica brasileira é propagada em inúmeros países, possibilitando visibilidade dos trabalhos em outros países. Além dessa distribuição por países, também foi possível identificar a sua distribuição por estados da federação.

Classificação	Estado	Total
1	São Paulo	25,36%
2	Minas Gerais	11,62%
3	Rio Grande do Sul	9,88%
4	Paraná	9,83%
5	Rio de Janeiro	7,78%
6	Santa Catarina	4,91%
7	Bahia	4,57%
8	Ceará	3,95%
9	Pernambuco	3,81%
10	Rio Grande do Norte	2,64%
11	Paraíba	2,50%
12	Goiás	2,14%
13	Distrito Federal	2,11%
14	Pará	1,48%
15	Espírito Santo	1,38%
16	Mato Grosso do Sul	1,17%

17	Mato Grosso	1,00%
18	Alagoas	0,94%
19	Sergipe	0,93%
20	Amazonas	0,64%
21	Maranhão	0,51%
22	Piauí	0,46%
23	Tocantins	0,19%
24	Acre	0,08%
25	Rondônia	0,05%
26	Roraima	0,05%
27	Amapá	0,03%

Tabela 1 – Eventos por estados da federação.
Fonte: Elaborado pelo autor (2022).

Conforme pode ser observado, os dois estados que possuem maior concentração de eventos estão localizados na região Sudeste: São Paulo e Minas Gerais. Seguidamente de estados na região Sul: Rio Grande do Sul e Paraná. Já os 3 estados que apresentaram menor volume de eventos foram: Roraima, Rondônia e Amapá. A distribuição apresentada pode estar relacionada ao fato de que a maior parte dos estados estão localizados na região Sudeste e Sul, onde estão as maiores instituições de ensino do país. De acordo com Dias (2016) a região Sudeste- Sul concentra 62,6% dos indivíduos com endereço profissional cadastrado em seus currículos. Em vista disso, faz sentido que a maioria dos eventos sejam nessas regiões, propiciando assim, maior facilidade da participação dos indivíduos. Além da maior concentração de indivíduos, essas regiões também podem estar relacionadas à concentração das principais instituições de ensino públicas do país, onde se agrupam os principais cursos de pós-graduação.

Partindo para uma análise mais específica a respeito da regionalidade dos eventos, foi realizada a caracterização dos eventos por cidades, totalizando cerca de 6.000 nomes de cidades encontradas após curadoria de dados. Também pôde-se verificar que existem 1.217.771 trabalhos publicados em anais de eventos não possuem cidade vinculada. A Tabela 2 apresenta a classificação das 30 cidades que possuem maior volume de trabalhos em eventos.

Classificação	UF	Cidade	Total
1	SP	São Paulo	630.885
2	RJ	Rio De Janeiro	479.965
3	CE	Fortaleza	304.453
4	BA	Salvador	282.003
5	SC	Florianópolis	281.386
6	RS	Porto Alegre	269.656
7	PR	Curitiba	260.442
8	MG	Belo Horizonte	256.105
9	SP	Águas De Lindóia	246.934
10	SP	Campinas	229.971
11	RN	Natal	199.188
12	PR	Foz Do Iguaçu	183.014
13	DF	Brasília	173.070
14	MG	Caxambu	160.960
15	SP	Ribeirão Preto	145.843
16	PB	João Pessoa	143.686
17	GO	Goiânia	137.544
18	RS	Gramado	135.260
19	PA	Belém	120.145
20	PR	Londrina	111.026
21	RS	Santa Maria	108.941
22	MG	Uberlândia	99.975
23	PR	Maringá	98.163
24	SP	São Carlos	90.252
25	RS	Pelotas	85.645
26	ES	Vitoria	84.173
27	AL	Maceió	74.510
28	SP	Bauru	72.937
29	MG	Viçosa	68.569
30	MT	Cuiabá	66.653

Tabela 2 – Total de trabalhos por cidade.
Fonte: Elaborado pelo autor (2022).

Dentre as 30 cidades que possuem mais trabalhos publicados em eventos é perceptível a diferença nos totais, a diferença para a 1ª posição (São Paulo) e a 2ª posição (Rio de Janeiro) é de 150.920 trabalhos. Ao se comparar os resultados apresentados na Tabela 1, percebe-se que houve algumas diferenças entre as duas cidades do Nordeste entre as 5 primeiras classificadas: Fortaleza, na posição 3 e Salvador, na posição 4.

Ou seja, não são cidades da região Sudeste-Sul, podendo assim, serem regiões que possam ganhar destaques futuramente, hipoteticamente por sua localização geográfica estar em uma área litorânea. Outra característica é que 14 cidades são capitais dos seus respectivos estados, demonstrando assim a importância de os eventos serem em cidades com maior

facilidade de acesso, mediante a questões de transporte como aeroportos.

4 Considerações Finais

A partir dos resultados e análises aqui apresentados, observou-se que os dados extraídos da Plataforma Lattes são uma excelente fonte para compreender como acontece a produção científica brasileira em anais de congresso, visto ser complexo encontrar um repositório com todos os dados de eventos científicos. Através das análises realizadas pode-se observar algumas características gerais no que tange a regionalidade dos eventos, sendo um fator importante, possibilitando compreender onde ocorrem os eventos nacionais e internacionais. Quais são as regiões do Brasil que possuem maior representatividade de eventos e cidades. Sendo possível constatar que existem regiões e cidades predominantes.

Importante destacar que devido ao preenchimento manual realizado pelos indivíduos no ato de atualização de um currículo, diversos erros, principalmente de digitação em geral leva a não consideração de um determinado registro.

Referências

CARMONA, Ingrid Valadares; PEREIRA, Vinicius Pereira. Ciência, tecnologia e sociedade e educação ambiental: uma revisão bibliográfica em anais de eventos científicos da área de ensino de ciências. Revista Ciências & Ideias, v. 8, n. 3, p. 94–114, 2018.

DIAS, Thiago Magela. Rodrigues. Um Estudo da Produção Científica Brasileira a partir de Dados da Plataforma Lattes. 181 p. Tese (Doutorado em Modelagem Matemática e Computacional) — Centro Federal de Educação Tecnológica de Minas Gerais, Setembro 2016.

DOMINGUES, I. O sistema de comunicação da ciência e o taylorismo acadêmico: questionamentos e alternativas. Estudos avançados, SciELO Brasil, v. 28, n. 82, p. 225–250, 2014.

PLATAFORMA LATTES. História do surgimento da Plataforma Lattes. Brasil, 2022.

SERRA, F. A. R.; FIATES, G. G.; FERREIRA, M. P. Publicar é difícil ou faltam competências? o desafio de pesquisar e publicar em revistas científicas na visão de editores e revisores internacionais. RAM. Revista de Administração Mackenzie, SciELO Brasil, v. 9, n. 4, p. 32–55, 2008.

Apêndice 01

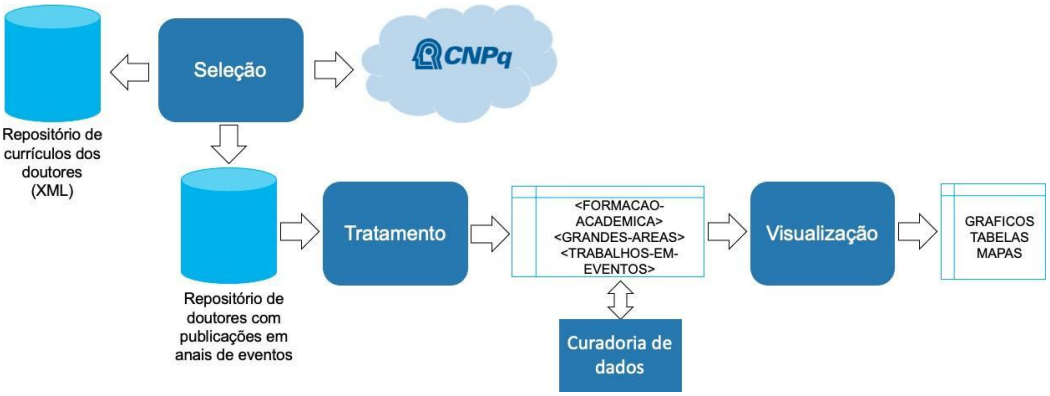


Figura 1 – Processo de seleção, tratamento e visualização dos dados.
Fonte: Elaborado pelo autor (2022).

Apêndice 02

Figura 2 – Ranking - Países com maior publicação de artigos em anais de eventos.
Fonte: Elaborado pelo autor (2022).

BRASILIANA MUSEUS: TESTE FUNCIONAL DO AGREGADOR DE DADOS MUSEAIS DO INSTITUTO BRASILEIRO DE MUSEUS

Brasiliana Museums: functional test of the museum data aggregator of the Brazilian Institute of Museums

Joyce Siqueira[1], Dalton Lopes Martins[2], Vinícius Nunes Medeiros[3]
(1) Universidade de Brasília, Brasília-DF, joycitta@gmail.com
(2) Universidade de Brasília, Brasília-DF, daltonmartins@unb.br
(3) Tainacan.Org, Brasília-DF, vnicius.nm.ba@gmail.com

Resumo
Os museus vinculados ao Instituto Brasileiro de Museus utilizam o software Tainacan para gerir e publicar seus acervos na Web. Assim, para facilitar a busca e recuperação da informação, um novo serviço foi desenvolvido, o Brasiliana Museus, um portal de acervos agregados que oferece uma interface única de busca e recuperação de dados na Web. O Portal foi recentemente desenvolvido e, no momento, encontra-se em fase de testes. Dessa forma, o presente artigo demonstra e discute os resultados obtidos por meio do teste de software intitulado teste funcional ou teste de caixa-preta, que visa validar funcionalidades, além de identificar possíveis falhas. Para tal, sete casos de teste foram elencados, são eles: inserir novo item; inserir novo item com termo de tesauro inexistente; atualizar item com alteração no conteúdo de um metadado; atualizar o nome da coleção; alterar a URL da instalação; excluir um item e; excluir um item, a partir da alteração do status de "público" para "privado". Os resultados apresentam que melhorias importantes devem ser implementadas, principalmente no âmbito da exclusão de itens, que obteve resultado insatisfatório em todos os testes. No entanto, também foi possível observar que a inclusão e edição itens funciona corretamente, além de identificar e corrigir problemas, aprimorando o serviço. Destaca-se que embora os testes tenham sido realizados no contexto do Brasiliana Museus, os resultados podem auxiliar a qualquer outra aplicação da solução de agregação desenvolvida.

Palavras-chave: Agregador; Brasiliana Museus; Tainacan; Teste Funcional; Teste de Software.

Abstract
Museums linked to the Brazilian Institute of Museums use the Tainacan software to manage and publish their collections on the Web. Thus, to facilitate the search and retrieval of information, a new service was developed, Brasiliana Museums, a portal of aggregated collections that offers a unique interface for searching and retrieving data on the Web. The Portal was recently developed and is currently in the testing phase. In this way, this article demonstrates and discusses the results obtained through the software test called functional test or black box test, which aims to validate functionalities, in addition to identifying possible failures. To this end, seven test cases were listed, they are: insert new item; insert new item with non-existent thesaurus term; update item with change in metadata content; update the collection name; change the installation URL; delete an item and delete an item, from status change from "public" to "private". The results show that important improvements must be implemented, mainly in the scope of the exclusion of items, which obtained unsatisfactory results in all tests. However, it was also possible to observe that adding and editing items works correctly, beyond that identifying and correcting problems, improving the service. It is noteworthy that although the tests were carried out in the context of Brasiliana Museums, the results can help any other application of the aggregation solution developed.

Keywords: Aggregator; Brasiliana Museums; Tainacan; Functional test; Software Testing.

1 Introdução

O Instituto Brasileiro de Museus – Ibram, administra diretamente 30 museus federais. Destes, 20 utilizam o *software* livre e de código aberto Tainacan (MARTINS; MARTINS, 2021), uma ferramenta flexível e robusta para WordPress que permite a gestão e a publicação de coleções digitais (TAINACAN.ORG, 2022), totalizando 22 coleções e mais de 17 mil itens na Web. Em franca expansão, o intuito é migrar para o digital todos os museus federais vinculados.

A fim de alcançar maior eficiência na busca e recuperação destes itens, beneficiando tanto usuários leigos quanto especialistas em documentação, o Ibram, em parceria com a Universidade de Brasília, desenvolveu o Portal Brasiliana Museus, um

serviço de agregação automatizado de acervos, que os integra em uma única instalação Tainacan (SIQUEIRA; MARTINS; MEDEIROS, 2022a; SIQUEIRA; MARTINS; LEMOS, 2022).

O Portal está disponível on-line, porém, dado seu recente desenvolvimento, ainda se encontra em fase de testes, necessários para a qualidade do serviço. Neste contexto, esse artigo apresenta os resultados obtidos por meio da aplicação de uma técnica de teste de software, intitulado teste funcional ou, como também é conhecido, teste de caixa-preta, visando auxiliar na validação das principais funcionalidades da aplicação.

Segundo Lamounier (2021), o teste funcional independe do conhecimento do comportamento interno do software, mas sim, de suas funcionalidades. Dessa forma, para o teste funcional são definidas especificações a partir dos requisitos funcionais do sistema, que são testadas para validar se as respostas ocorrem de acordo com o esperado. Com isso, objetiva identificar anomalias preocupando-se apenas com critérios fundamentais dos sistemas.

Foram elencadas sete especificações, que aqui denominamos de casos de teste, voltados à análise do comportamento do sistema durante a coleta dos dados modificados pelos museus na base de dados de origem.

Assim, o estudo se divide em: seção 2, Objetivo, que destaca o objetivo geral da pesquisa; seção 3, Portal Brasiliana Museus, que apresente e projeto e trata de especificidades relevantes a compreensão dos casos de teste escolhidos; seção 4, Procedimentos metodológicos; seção 5, Resultados, onde são apresentados os comportamentos do sistema em cada caso de testes, assim como observações e correções, quando possível; e, por fim, a seção 6, tece as considerações finais.

2 Objetivo Geral

Aplicar a técnica de teste de software, intitulado teste funcional ou teste de caixa-preta, utilizando o critério de particionamento de equivalência, no serviço de agregação de dados do Portal Brasiliana Museus, a fim de validar suas funcionalidades e identificar possíveis problemas.

3 Portal Brasiliana Museus

O Portal Brasiliana Museus (Acervo em Rede[1]), disponível em http://integracaoibram.tainacan.org/, é um portal agregador de objetos digitais museais do Ibram, que permite a busca e a recuperação dos acervos em uma interface única. Utiliza o software Tainacan e, por isso, além do repositório digital, conta com recursos de criação de páginas do WordPress (SIQUEIRA; MARTINS; MEDEIROS, 2022a; SIQUEIRA; MARTINS; LEMOS, 2022). A Figura 01 apresenta a página principal do Portal.

Figura 01. Portal Brasiliana Museus

Fonte: elaborada pelos autores (2022)

Os acervos museais do Ibram têm os metadados dos itens coletados, agregados, transformados e armazenados, via plugin de submissão, de forma automatizada, em uma nova instalação Tainacan por meio das ferramentas do Elastic Stack (SIQUEIRA; MARTINS; MEDEIROS, 2022a). A arquitetura simplificada do serviço de agregação está apresentada na Figura 02.

Figura 02. Arquitetura simplificada

Fonte: elaborada pelos autores (2022)

De acordo com a wiki do Tainacan (2022), os itens são organizados por coleções, que são configuradas conforme as características dos itens que farão parte dela.

[1] A interface do Portal ainda se mantém com o nome do projeto, intitulado "Acervo em Rede".

Os usuários são livres para modelar quaisquer metadados, além determinar se estes são públicos (visíveis por todos) ou privados (visíveis apenas para editores da coleção).

O Ibram modela seus metadados a partir do Inventário Nacional dos Bens Culturais Musealizados - INBCM, composto por 15 elementos descritivos que devem ser utilizados por todos os museus federais (BRASIL, 2021).

Assim, no processo de agregação apenas os metadados públicos mapeados para o INBCM são coletados (SIQUEIRA; MARTINS; MEDEIROS, 2022b). Um deles, o elemento descritivo "classificação", deve ser preenchido por uma linguagem documentária, a saber: o Tesauro de Objetos do Patrimônio Cultural nos Museus Brasileiros (FERREZ, 2016).

Na seção 4, são descritos os procedimentos metodológicos utilizados para realização dos testes.

4 Procedimentos Metodológicos

A presente pesquisa classifica-se como sendo de natureza qualitativa e descritiva; como procedimento técnico utiliza-se a pesquisa bibliográfica; como método utiliza-se técnica específica da área de Engenharia de Software, o teste funcional ou teste de caixa-preta com o critério de particionamento de equivalência, do serviço de busca e recuperação da informação agregada dos acervos digitais do Instituto Brasileiro de Museus, intitulado Brasiliana Museus.

No teste funcional o componente de software é abordado como se fosse uma caixa-preta, ou seja, não se considera o comportamento interno do código-fonte, mas sim se as saídas esperadas são obtidas por meio das entradas de dados testadas (LAMOUNIER, 2021; PRESSMAN; MAXIM, 2021).

Alguns critérios podem ser adotados para os testes funcionais. Aqui optou-se pelo critério de particionamento por equivalência, que tem como finalidade tornar a quantidade de dados a serem testadas finitas. Este critério divide a entrada em classes por equivalência, podendo-se assumir que qualquer elemento da classe pode ser considerado um representante fiel dela. Suas classes são definidas por um conjunto de classes válidas e inválidas (LAMOUNIER, 2021; PRESSMAN; MAXIM, 2021).

A arquitetura demonstrada na Figura 2 foi inicialmente validada por meio do próprio Portal Brasiliana Museus, que demonstra o sucesso da coleta e da integração dos dados em uma nova instalação Tainacan. No entanto, as bases de origem, ou seja, as bases dos Museus podem, a qualquer momento, serem atualizadas e uma nova coleta de dados deve ser realizada.

Nesse sentido, esta pesquisa foca em validar a atualização dos dados no Portal Brasiliana Museus a partir da atualização nos dados de origem, realizando os casos de teste, do Quadro 01.

Quadro 01. Casos de testes

1. inserir novo item
2. inserir novo item com termo de tesauro inexistente
3. atualizar item com alteração no conteúdo de um metadado
4. atualizar o nome da coleção
5. atualizar a URL da instalação de origem
6. excluir um item
7. excluir um item, a partir da alteração do status de "público" para "privado"

Fonte: elaborado pelos autores (2022)

Há certamente mais de sete casos de testes possíveis a serem testados, no entanto, metodologicamente buscou-se casos de teste com funcionalidades essenciais, para verificar seu funcionamento, e casos de testes previamente detectados pelos desenvolvedores como sendo críticos e passíveis de problemas.

4.1 Descrição dos casos de teste

Os casos descritos no Quadro 01 estão detalhados a seguir.

Caso de teste 01 – inserir novo item: **descrição:** um novo item é incluído na base de dados de origem e uma nova coleta de dados é realizada; **pré-requisito:** o novo item deve possuir elemento descritivo "classificação" preenchido com uma taxonomia compatível com o Tesauro de Objetos do Patrimônio Cultural nos Museus Brasileiros; **passos do teste:** 1. incluir um novo item na base de dados de origem e; 2. realizar nova coleta de dados; **resultado esperado:** um novo item deve ser inserido na coleção no Brasiliana Museus.

Caso de teste 02 – inserir novo item com termo de tesauro inexistente: **descrição**: um novo item é incluído na base de dados de origem e uma nova coleta de dados é realizada; **pré-requisito**: o novo item deve conter elemento descritivo “classificação” preenchido com termo de tesauro inexistente no Tesauro de Objetos do Patrimônio Cultural nos Museus Brasileiros; **passos do teste:** 1. incluir um novo item na base de dados de origem e; 2. realizar nova coleta de dados; **resultado esperado**: o novo item não deve ser inserido na coleção no Brasiliana Museus.

Caso de teste 03 – atualizar item com alteração no conteúdo de um metadado: **descrição:** um item terá um dos metadados editados na base de dados de origem e uma nova coleta de dados é realizada; **pré-requisito**: o item deve possuir elemento descritivo “classificação” preenchido com uma taxonomia compatível com o Tesauro de Objetos do Patrimônio Cultural nos Museus Brasileiros; **passos do teste:** 1. editar o conteúdo de qualquer metadado do item na base de dados de origem e; 2. realizar nova coleta de dados; **resultado esperado:** o item deve ter o conteúdo o metadado atualizado na coleção, sem resultar em duplicidade.

Caso de teste 04 – atualizar o nome da coleção: **descrição:** o nome da coleção é editado na base de dados de origem e uma nova coleta de dados é realizada; **pré-requisito:** nenhum; **passos do teste:** 1. alterar o nome da coleção na base de dados de origem e; 2. realizar nova coleta de dados; **resultado esperado:** o nome da coleção deve ser atualizado, sem duplicidade e sem quebra de links.

Caso de teste 05 – atualizar a URL (*Uniform Resource Locator*) da instalação de origem: **descrição:** a URL da instalação de origem é modificada e uma nova coleta de dados é realizada; **pré-requisito:** possuir uma URL válida; **passos do teste:** 1. alterar a URL da instalação de origem e; 2. realizar nova coleta de dados; **resultado esperado:** a coleta deve ser realizada normalmente, sem resultar em erros, duplicação de itens e quebra de links.

Caso de teste 06 – excluir um item: **descrição:** excluir um item na base de dados de origem e realizar uma nova coleta de dados; **pré-requisito:** nenhum; **passos do teste:** 1. excluir um item na base de dados de origem e; 2. realizar nova coleta de dados; **resultado esperado**: o item deve ser excluído no Brasiliana Museus.

Caso de teste 07 – excluir um item, a partir da alteração do status de “público” para “privado”: **descrição:** apenas itens públicos são coletados, assim, caso o item seja alterado para privado na base de origem, o mesmo deve ser excluído do Brasiliana Museus, após a realização de uma nova coleta de dados; **pré-requisito:** nenhum; **passos do teste:** 1. alterar um item de "público" para “privado” na base de dados de origem e; 2. realizar nova coleta de dados; **resultado esperado:** o item deve ser excluído no Brasiliana Museus.

Os casos de testes foram aplicados em ambiente real de produção, a partir da coleta de dados dos museus vinculados e sua integração no Portal Brasiliana Museus. Os resultados estão apresentados na seção 5.

5 Resultados

Os sete casos de teste foram realizados, obtendo os resultados descritos. Em complemento são descritas observações pertinentes a cada resultado, além de possíveis correções a serem realizadas para versão futura do Brasiliana Museus.

Caso de teste 1 – inserir novo item: **resultado:** o novo item é corretamente coletado e inserido na base de dados do Brasiliana Museus; **observação**: não há; **sugestão de correção**: não há.

Caso de teste 2 – inserir novo item com termo de tesauro inexistente: **resultado:** o novo item é coletado, porém não é inserido na base de dados do Brasiliana Museus. O item coletado é enviado para um local de armazenamento desenvolvido especificamente para receber itens coletados com erros; **observação:** o item não é armazenado no Brasiliana Museus porque o Tainacan está configurado corretamente para não permitir a criação de novos termos. Caso o Ibram opte por aceitar termos de tesauro diferentes, basta configurar o Tainacan para tal; **sugestão de correção:** não há.

Caso de teste 3 – atualizar item com alteração no conteúdo de um metadado: **resultado:** o item é corretamente coletado e atualizado na base de dados do Brasiliana Museus, não causando duplicação;

observação: não há; **sugestão de correção:** não há.

Caso de teste 4 – atualizar o nome da coleção: **resultado:** não ocorre duplicação e/ou quebra de links na base de dados do Brasiliana Museus; **observação:** a coleta dos dados considera o código da coleção e não seu nome. No entanto, em cada item do Brasiliana Museus é inserido manualmente, via pipeline do Filebeat, o metadado "coleção", que recebe o nome dela. Dessa maneira, a nova coleta não atualiza esse metadado; **sugestão de correção:** o pipeline do Filebeat deve ser atualizado, contendo o novo nome da coleção, e uma nova coleta realizada.

Caso de teste 5 – atualizar a URL da instalação de origem: **resultado:** os itens são coletados, porém, por possuírem URL diferente dos da primeira coleta, todas as coleções existentes são duplicadas no Brasiliana Museus, como se fossem duas fontes de origem distintas. Além disso, há quebra dos links da coleção anteriormente coletada; **observação:** a URL é um dos dados utilizados para criar o identificador único de cada item no Brasiliana Museus. Dessa forma, com a mudança da URL um novo identificador é gerado e, por isso, os itens são duplicados. Nesse caso, a solução imediata é a exclusão das coleções anteriores a alteração da URL; **sugestão de correção:** alterar a configuração do pipeline do Logstash para que não seja utilizada a URL na criação do identificador único, assim como, nenhum outro dado que possa ser modificado na base de origem.

Caso de teste 6 – excluir um item: **resultado:** os itens não são coletados, porém, também não são excluídos da base de dados do Brasiliana Museus; **observação:** não há implementada nenhuma funcionalidade que verifique se um item foi excluído, assim, mesmo que ele não seja coletado, ele continua na base de dados. Nesse caso, a solução imediata é a exclusão de toda base de dados e uma nova coleta; **sugestão de correção:** não há uma solução a ser apresentada. Será necessário a realização de estudos aprofundados.

Caso de teste 7 – excluir um item, a partir da alteração do status de "público" para "privado": **resultado:** os itens não são coletados, porém, também não são excluídos da base de dados do Brasiliana Museus; **observação:** embora tenha uma diferente motivação, este caso de teste deveria realizar a mesma ação do caso de teste 6, ou seja, deveria excluir um item da base. O que também não ocorre, sendo necessária a exclusão da base e uma nova coleta de dados.

De forma simplificada os testes apresentaram o resultado demonstrado no Quadro 02.

Quadro 02. Resultado simplificado dos testes

N.	Caso de Teste	Resultado
1	inserir novo item	OK
2	inserir novo item com termo de tesauro inexistente	OK
3	atualizar item com alteração no conteúdo de um metadado	OK
4	atualizar o nome da coleção	OK
5	atualizar a URL da instalação de origem	X
6	excluir um item	X
7	excluir um item, a partir da alteração do status de "público" para "privado"	X

Fonte: elaborado pelos autores (2022)

Os resultados apontam que a inserção e atualização de dados na base agregada acontece da forma correta, e que atualizar metadados que são gerados manualmente via pipeline do Filebeat sempre requerem atenção especial, que é o caso do nome da coleção (Quadro 2 - n. 4).

Por outro lado, um tópico essencial, que é excluir um item da base de dados agregada, não funcionou em nenhum dos casos de teste. Sendo necessário apagar toda a base e recomeçar o processo de coleta, o que, embora resolva a primeiro momento, não é o ideal. Assim, constatou-se que a exclusão é um processo complexo, que requer maiores pesquisas para ser implementado de forma satisfatória.

O caso de teste 5, atualizar a URL da instalação, apresentou uma fragilidade importante na regra de negócio usada para gerar os identificadores únicos dos itens no agregador, ou seja, esse tipo de código não pode ser gerado a partir de dados que podem ser modificados nas bases de origem.

O resultado apresenta importantes fragilidades na solução em relação à fase de atualização dos dados previamente coletados, abrindo oportunidade para novos estudos visando aprimorar requisitos tão essenciais à construção de um agregador de dados.

6 Considerações Finais

O Portal sem dúvidas representa um significativo avanço para área, pela incipiência do tema no Brasil em contraste com países como Estados Unidos da América e Europa (SIQUEIRA; MARTINS, 2021)

Pode-se explorar as funcionalidades do serviço de agregação a partir da ótica da atualização dos dados coletados, identificando resultados positivos na inclusão e alteração de itens; resultados negativos na exclusão de itens e; alguns pontos de atenção, como a mudança de dados incluídos manualmente e o uso de dados que podem ser modificados na origem para criação de identificadores.

É importante destacar que o serviço de agregação pode ser implantado em outros contextos, e estes resultados podem ser considerados para qualquer outra aplicação.

Referências

BRASIL. Instituto Brasileiro de Museus, Resolução Normativa Ibram nº 6, de 31 de agosto de 2021. Diário Oficial. Disponível em: https://www.in.gov.br/Web/dou/-/resolucao-normativa-ibram-n-6-de-31-de-agosto-de-2021-342359740. Acesso em 17 set. 2022.

FERREZ, Helena Dodd. Tesauro de objetos do patrimônio cultural nos museus brasileiros. Rio de Janeiro: Fazer Arte. Gerência de Museus da Secretaria Municipal de Cultura, 2016. Disponível em: https://tesauromuseus.com.br/download/tesauro.pdf. Acesso em 17 set. 2022.

LAMOUNIER, Stella Marys Dornelas. Teste e inspeção de software, técnicas e automatização. São Paulo. Conteúdo Saraiva. 2021-1, recurso online.

MARTINS, Dalton Lopes; MARTINS, Luciana Conrado. Desafios e Aprendizados na Implantação do Tainacan nos Museus do Instituto Brasileiro de Museus. Revista Eletrônica Ventilando Acervos - Florianópolis: MVM, 2021 – ISSN 2318-6062. Disponível em: http://ventilandoacervos.museus.gov.br . Acesso em 15 set. 2022

PRESSMAN, Roger S.; MAXIM, Bruce R. Engenharia de software. Editora McGrawHill: 9. ed. – Porto Alegre: AMGH, 2021. E-pub.

SIQUEIRA, Joyce; MARTINS, Dalton Lopes. Workflow models for aggregating cultural heritage data on the Web: A systematic literature review. Journal of the Association for Information Science and Technology, 2021. DOI: https://doi.org/10.1002/asi.24498

SIQUEIRA, Joyce; MARTINS, Dalton Lopes; LEMOS, Daniela Lucas da Silva. Brasiliana Museus: Serviço de Busca e Recuperação da Informação Agregada dos Acervos Digitais do Instituto Brasileiro de Museus. ENANCIB, 2022.

SIQUEIRA, Joyce; MARTINS, Dalton Lopes; MEDEIROS, Vinicius Nunes. Workflow with emphasis in technologies for the construction of an aggregator and a dashboard of Brazilian museums digital collections. 2022a. 3rd EAI International Conference on Data and Information in Online Environments.

SIQUEIRA, Joyce; MARTINS, Dalton Lopes; MEDEIROS, Vinícius Nunes. Metadata mapping in Tainacan: new functionality for digital museums linked to the Brazilian Institute of Museums. Advanced Notes in Information Science, v. 2, p. 182-191, 2022b. DOI: https://doi.org/10.47909/anis.978-9916-9760-3-6.92

TAINACAN. Wiki Tainacan - Itens, 2022. Disponível em: https://tainacan.github.io/tainacan-wiki/#/pt-br/items?id=criar-itens. Acesso em 16 set. 2022.

TAINACAN.ORG. Página Principal, 2022. Disponível em: https://tainacan.org/ Acesso em 14 set. 2022.

CIÊNCIA DE DADOS: UMA REVISÃO DE LITERATURA EM PERIÓDICOS DA CIÊNCIA DA INFORMAÇÃO

DATA SCIENCE: A LITERATURE REVIEW IN INFORMATION SCIENCE JOURNALS

Aurea Celeste Pires de Souza[1]
Clarice Luzia Casoni[2]
Merabe Carvalho Ferreira da Gama[3]
José Eduardo Santarem Segundo[4]
(1) Universidade Estadual de Londrina (UEL), aureaceleste.souza@gmail.com
(2) Universidade Estadual de Londrina (UEL), claricecasoni@gmail.com
(3) Universidade Estadual de Londrina (UEL), merabecarvalho@yahoo.com.br
(4) Universidade de São Paulo (USP), santarem@usp.br

Resumo
Tem-se como **proposta** identificar pesquisas que abordem a temática de Ciência de Dados em periódicos de estratos mais elevados da Ciência da Informação. O **objetivo da pesquisa** consiste em identificar a produção científica sobre Ciência de Dados em periódicos brasileiros em Ciência da Informação. O **método** utilizado neste estudo foi a Revisão Sistemática da Literatura (RSL) de estudos que abordem o tema Ciência de Dados, em suas variadas interpelações, a partir de periódicos científicos da Ciência da Informação, de estratos Qualis A1 e A2, no período de 2018 a 2021. **Resultados:** Após a seleção dos periódicos da área da Ciência da Informação com estratos Qualis A1 e A2, foram selecionados 40 artigos em 05 periódicos com produções relacionadas à temática, os quais foram categorizados em 15 categorias de assunto. **Conclusões**: Na análise percebeu-se que o ano de 2018 foi o mais produtivo. Além disso, os pesquisadores têm dado preferência por publicar em periódicos Qualis A2. Três categorias se destacaram com maior frequência: Big Data (25,5%), Dados de Pesquisa (16,2%) e Dados Abertos (13,9%).

Palavras-chave: Ciência de Dados; Ciência da Informação; Produção Científica; Estrato Qualis.

Abstract
The proposal is to identify researches that approach the theme of Data Science in journals of higher strata of Information Science. The method used in this study was the Systematic Literature Review (RSL). The objective of the research is to carry out a Systematic Literature Review (RSL) of studies that address the topic of Data Science, in its various interpellations, from scientific journals of Information Science, from Qualis A1 and A2 strata, in the period of 2018 to 2021. Results: After the selection of journals in the area of Information Science with Qualis A1 and A2 strata, 40 articles were selected in 05 journals with productions related to the theme, which were categorized into 15 subject categories. Conclusions: In the analysis, it was noticed that the year 2018 was the most productive. In addition, researchers have preferred to publish in Qualis A2 journals. Three categories stood out most frequently: Big Data (25.5%), Research Data (16.2.2%) and Open Data (13.9%).

Keywords: Data Science; Information Science; Scientific production; Qualis Stratum.

1 Introdução

Com o grande volume de dados produzidos continuamente nas plataformas digitais em todos os aspectos, a nível organizacional, institucional e governamental, tanto no âmbito acadêmico quanto profissional, faz-se necessário que esses dados sejam tratados e dispostos de forma que permita ao observador ter perspectivas futuras e condições de juízo de valor quanto ao presente.

Na Ciência da Informação, dentre as definições apresentadas, “dado” é “[...] um elemento ‘passível de ser compreendido, interpretado, comunicado e processado’, indica uma expectativa de mudança de estado entre as etapas de apreensão, tratamento e, supostamente, resultado.” (SEMIDÃO, 2014, p. 96). Neste contexto, a Ciência de Dados está incumbida da obtenção de dados gerados em fontes variadas e veiculadas na web e padronizar o modo de tratar os dados e metadados, de forma a permitir que sejam explorados e resultem em informação estratégica na tomada de decisão (RAUTENBERG; CARMO, 2019).

No cenário atual, a Ciência de Dados está no centro dessa questão e possui relação com a Ciência da Informação. Portanto, esta pesquisa parte da seguinte questão: Qual a produção científica sobre Ciência de Dados nos periódicos brasileiros de Ciência da informação de estratos Qualis A1 e A2?

2 Objetivos

Identificar a produção científica sobre Ciência de Dados em periódicos brasileiros em Ciência da Informação.

3 Procedimentos Metodológicos

Esta pesquisa é do tipo bibliográfica e utilizou a técnica de Revisão Sistemática de Literatura. Optou-se por este tipo de técnica pois ela possui potencial para encontrar, sintetizar e avaliar diversos estudos sobre um tema, proporcionando um caráter mais fidedigno à revisão de literatura (TRANFIELD; DENYER; SMART, 2003).

A pesquisa iniciou com a definição do Protocolo de busca (Apêndice A), a primeira etapa de uma Revisão Sistemática de Literatura, e que visa proporcionar uma fundamentação lógica para a realização do processo de coleta e análise dos dados.

O protocolo (Apêndice A) previu as fontes de informação, as estratégias de busca, o recorte temporal, os critérios de inclusão e exclusão, dentre outros aspectos.

Após a definição do protocolo de busca, consultou-se a Plataforma Sucupira, a fim de identificar os periódicos A1 e A2 na área de Comunicação e Informação. Após essa identificação foram verificados o foco e escopo de cada periódico para selecionar apenas aqueles que são inerentes à Ciência da Informação.

Foram identificados seis periódicos que atenderam a esse propósito: três de Qualis A1 (Informação & Sociedade, Perspectiva em Ciência da Informação e Transinformação) e três de Qualis A2 (Em Questão, Encontros Bibli e Informação & Informação).

Procedeu-se então às pesquisas em cada um dos periódicos identificados, conforme as estratégias e campos de busca previstos no protocolo (Apêndice A), tendo como recorte temporal o período de 2018 a 2021. Não foi possível realizar a busca no periódico Informação & Informação pois a plataforma encontrava-se em manutenção. As buscas foram realizadas no mês de agosto de 2022.

Recuperou-se ao total 1.913 artigos nos periódicos selecionados, utilizando seis estratégias de busca. Observou-se que muitos artigos se repetiam em todas as estratégias de busca, o que gerou a exclusão de artigos duplicados e não pertinentes à pesquisa. Após uma análise preliminar dos títulos, resumos, palavras-chave e exclusão dos duplicados foram selecionados 40 artigos que apresentavam potencial para responder à questão norteadora da pesquisa (Apêndice C).

Procedeu-se à leitura de cada obra selecionada, utilizando como instrumento um roteiro de observação que buscou identificar: título do artigo, autor, ano de publicação e objetivo.

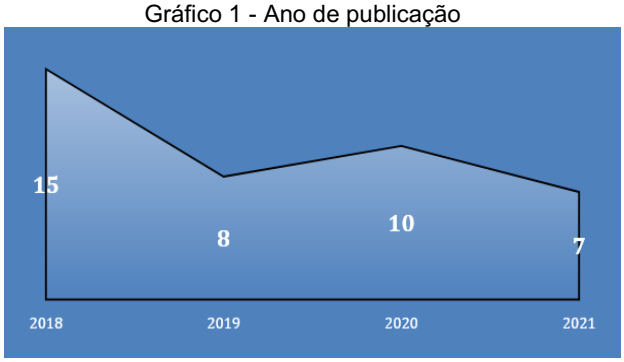
O tratamento dos dados ocorreu por meio de estatística descritiva e da análise de conteúdo do tipo categorial, conforme as recomendações de Bardin (2015). Para compor as categorias, foram analisados os títulos e objetivos dos artigos para definir um termo que representasse o assunto o qual o artigo tratava e assim, nomear a categoria. Ao final foram geradas 15 categorias (Apêndice B).

4 Resultados

Os resultados foram organizados de maneira qualitativa e quantitativa referente a quantidade de artigos por periódicos, ano de publicação e temática tratada.

O gráfico 1 apresenta a quantidade de artigos por ano de publicação.

Gráfico 1 - Ano de publicação



Fonte: Dados da Pesquisa, 2022.

Conforme os dados do gráfico 1, observa-se que o período no qual houve maior número de publicações a respeito da

temática, dentro do recorte da pesquisa, foi o ano de 2018, mantendo-se uma média em torno de 10 publicações nos anos posteriores.

A pesquisa não contempla o ano de 2022 para não se criar a falsa ideia de que a produção sobre o tema poderia ter declinado, visto que o ano de 2022 ainda está em curso e artigos submetidos e a submeter só poderão ser computados mais adiante.

Dados da plataforma Sucupira indicam que a Ciência da Informação brasileira conta com três periódicos de estrato Qualis A1 e três de estrato Qualis A2. O gráfico 2 expressa a quantidade de artigos por estrato Qualis no recorte da pesquisa.

Gráfico 2 - Estrato Qualis

Fonte: Dados da Pesquisa, 2022.

Dos artigos selecionados para a presente pesquisa, a maioria são de estrato Qualis A2. Cumpre ressaltar que a busca foi feita em três periódicos A1 e em dois periódicos A2, uma vez que a plataforma do periódico Informação & Informação durante a coleta de dados encontrava-se em manutenção.

Nestes termos, pode-se inferir que o número de artigos em estrato A2 pode ser ainda maior, revelando que os pesquisadores sobre a temática da Ciência de Dados, têm privilegiado periódicos A2 no momento de compartilharem os resultados de suas pesquisas.

O gráfico 3 demonstra os dados em relação a quantidade de artigos publicados em cada periódico pesquisado.

Gráfico 3 - Quantidade de artigos por Periódico

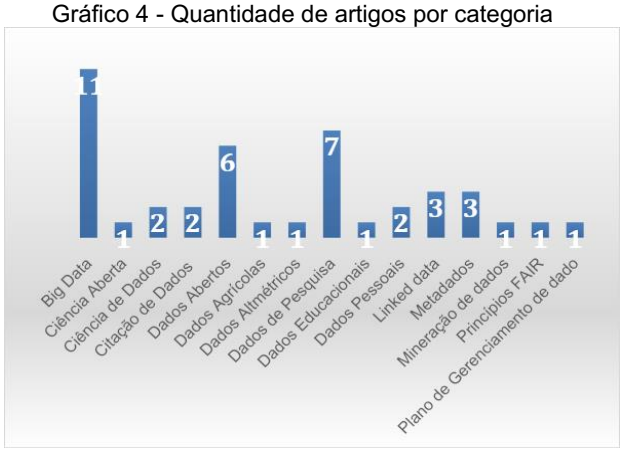
Fonte: Dados da Pesquisa, 2022.

Os dados do gráfico 3 indicam que o periódico Encontros Bibli, de estrato Qualis A2, foi o que mais publicou pesquisas a respeito da Ciência de Dados, de 2018 a 2021. Destaca-se que o número de artigos recuperados neste periódico é quase a metade de todos os artigos selecionados na pesquisa.

Sugere-se que pesquisas futuras de caráter explicativo sejam empreendidas a fim de identificar o porquê desses resultados.

No que se refere aos assuntos tratados nos artigos recuperados, foi possível organizá-los em 15 categorias, conforme expresso no gráfico 4. Procedeu-se à análise de conteúdo para identificar o tema principal que cada artigo tratava, a fim de categorizá-los. Em casos em que se observou mais de um tema predominante, incluiu-se o artigo em mais de uma categoria, o que ocorreu em três casos. Portanto, o somatório de todos os dados do gráfico 4 é 43.

Gráfico 4 - Quantidade de artigos por categoria



Fonte: Dados da Pesquisa, 2022.

O gráfico 4 permite constatar que a discussão sobre Ciência de Dados é ampla e com diversificadas aplicações. A maioria dos estudos publicados no escopo desta pesquisa estão relacionados a Big Data

(25,5%), Dados de Pesquisa (16,2%) e Dados Abertos (13,9%). Essas três categorias, nesta pesquisa, juntas, somam mais da metade (55,6%) dos estudos produzidos.

Entende-se que a aplicação da Ciência de Dados é percebida como "transformações profundas, no campo científico, que estão ocorrendo pela utilização de altos volumes de dados, permitindo a compreensão de diversos eventos que, até então, não haviam sido identificados." (CONEGLIAN; SANTAREM; SANT'ANA, 2017, p.62-63). Ressalta-se que, apesar de em baixa quantidade, outros assuntos relacionados a dados também estão presentes nesses periódicos, como: Mineração de dados, Plano de Gerenciamento de dados, dentre outros.

As palavras-chave dos artigos também foram analisadas, a partir das quais foi gerada uma nuvem de palavras (Apêndice D) no aplicativo Orange. Observou-se que as palavras-chave mais utilizadas pelos pesquisadores foram: Big data, Dados abertos, metadados, web semântica, recuperação da informação, dados científicos, ciência aberta e ciência da informação.

Durante a análise dos artigos foi possível identificar uma incipiência em relação à terminologia utilizada na indexação dos termos nas bases de dados, isto demanda novos estudos pois "a chamada era do Big Data trouxe grandes revoluções na maneira como os dados são analisados, tanto dentro da ciência, quanto nas empresas e até no governo."(CONEGLIAN; SANTARÉM; SANT'ANA, 2017, p.62-63).

5 Considerações Finais

Na análise percebeu-se que os pesquisadores têm publicado de forma mais consistente em periódicos Qualis A2. Três categorias se destacaram com maior frequência: Big Data (25,5%), Dados de Pesquisa (16,2%) e Dados Abertos (13,9%).

É importante ressaltar a intenção dessa pesquisa, de demonstrar que as principais revistas da área de Ciência da Informação no Brasil (estrato A1 e A2) têm recebido artigos de qualidade sobre o tema de Ciência de Dados, isso indica que pesquisas importantes sobre o tema têm sido realizadas no contexto da Ciência da Informação, revelando assim a possibilidade de surgirem novas pesquisas e um crescimento para o tema na Ciência da Informação.

Ao final desta revisão considera-se que o objetivo de identificar a produção científica sobre Ciência de Dados nestas fontes de informação foi alcançado e revelou dados que possibilitam a realização de pesquisas futuras.

Referências

BARDIN, Laurence. **Análise de Conteúdo.** São Paulo: Edições 70, 2015.

CONEGLIAN, Caio Saraiva; SANTAREM SEGUNDO, José Eduardo; SANT'ANA, Ricardo Cesar Gonçalves. Big Data: fatores potencialmente discriminatórios em análise de dados. **Em Questão**, Porto Alegre, v. 23, n. 1, p. 62–86, 2017. DOI: 10.19132/1808-5245231.62-86. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/62122. Acesso em: 7 set. 2022.

SEMIDÃO, Rafael Aparecido Moron. **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, Marília, 2014. Disponível em: https://repositorio.unesp.br/handle/11449/110783. Acesso em: 5 set. 2022.

RAUTENBERG, Sandro; CARMO, Paulo Ricardo Viviurka do. Big data e ciência de dados: complementaridade conceitual no processo de tomada de decisão. **Brazilian Journal of Information Science:** research trends, [*S.l.*], v. 13, n. 1, p. 56-67, 2019. Disponível em: https://revistas.marilia.unesp.br/index.php/bjis/article/view/8315. Acesso em: 5 set. 2022.

TRANFIELD, David; DENYER, David; SMART, Palminder. Towards a methodology for developing evidence- informed management knowledge by means of systematic review. **British Journal of Management**, [*S.l.*], v. 14, n. 3, p. 207-222, 2003.

Apêndice A – Protocolo da Revisão de Literatura Sistemática

Itens	Definições
Escopo do estudo	Produção científica sobre Data Science em periódicos da Ciência da Informação
Critérios de inclusão	Idioma: Português e Inglês Publicações: artigos publicados em periódicos nacionais de estrato Qualis A1 e A2. Pesquisas em andamento
Critérios de exclusão	Erratas, pesquisas duplicadas, fora do escopo do estudo ou que não oferecem texto completo. Tipologia: Editoriais, resenhas.
Periodicidade	2018 a 2021
Termos de busca	- Data Science - Ciência de dados - Ciência dos dados - Ciência de dados AND Ciência da Informação - Ciência dos dados AND Ciência da Informação - Data Science AND Ciência da Informação
Campos de busca	Todos
Campos de análise	Título, resumo, palavras-chave e texto completo
Fontes de informação	Periódicos de Ciência da Informação nacionais de estrato Qualis A1, A2 e B1 - Plataforma Sucupira (Periódicos) - Comunicação e Informação (Quadriênio 2013-2016)
Limitações de busca	Periódicos com instabilidade em seus sistemas

Fonte: Elaborado pelos autores (2022)

Apêndice B – Categorias geradas a partir da Análise de conteúdo categorial

Categorias geradas a partir da Análise de conteúdo categorial
Big Data Ciência Aberta Ciência de Dados Citação de Dados Dados Abertos Dados Agrícolas Dados Altmétricos Dados de Pesquisa Dados Educacionais Dados Pessoais Linked data Metadados Mineração de dados Princípios FAIR Plano de Gerenciamento de dados

Fonte: Elaborado pelos autores (2022)

Apêndice C – Lista dos artigos recuperados

Referências
ALVES, . Rosemari Pereira dos Santos; BORTOLIN, Sueli; ALCARÁ, Adriana Rosecler. Técnicas de análise de dados empregadas no Programa de Pós-graduação de Ciência da Informação da Universidade Estadual de Londrina. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 23, n. 51, p. 59–73, 2018. DOI: 10.5007/1518-2924.2018v23n51p59. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2018v23n51p59. Acesso em: 15 nov. 2022.

ARAÚJO, Ronaldo Ferreira de; APPEL, André Luiz. O financiamento de projetos de pesquisa em ciência aberta: análise de dados da base Dimensions. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 26, n. Especial, p. 1–19, 2021. DOI: 10.5007/1518-2924.2021.78828. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/78828. Acesso em: 15 nov. 2022.

BORBA, Vildeane da Rocha; CAREGNATO, Sônia Elisa. Agregadores de dados altmétricos: analisando o altmetric.com e o webometric analyst. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 26, n. Especial, p. 1–19, 2021. DOI: 10.5007/1518-2924.2021.78797. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/7879. Acesso em: 15 nov. 2022.

BRANDT, Mariana Baptista; VIDOTTI, Silvana Aparecida Borsetti Gregorio; SANTAREM SEGUNDO, José Eduardo. Modelo de dados abertos conectados para informação legislativa. **Informação & Sociedade: Estudos**, Belo Horizonte, v. 28, n. 2, 2018. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/37979. Acesso em: 15 nov. 2022.

BRANDT, Mariana Baptista; VIDOTTI, Silvana Aparecida Borsetti Gregório. Dados de pesquisa em informação legislativa. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 25, p. 1–14, 2020. DOI: 10.5007/1518-2924.2020.e72208. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2020.e72208. Acesso em: 15 nov. 2022.

CARVALHO, Marcelo Batista de; TSUNODA, Denise Fukumi. Análise de dados em artigos recuperados da Web of Science (WoS). **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, Edição Especial - v. 23, n. esp. 1, p. 112–125, 2018. DOI: 10.5007/1518-2924.2018v23nespp112. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2018v23nespp112 . Acesso em: 15 nov. 2022.

CONEGLIAN, Caio Saraiva; DIEGER, Rodrigo; SANTAREM SEGUNDO, José Eduardo; CAPRETZ, Miriam Akemi Manabe. O papel da web semântica nos processos do big data. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 23, n. 53, p. 137–146, 2018. DOI: 10.5007/1518-2924.2018v23n53p137. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2018v23n53p137. Acesso em: 15 nov. 2022.

CRISTOVÃO, Henrique Monteiro; FERNANDES, Jorge Henrique Cabral. Recuperação de informação em dados ligados: um modelo baseado em mapas conceituais e análise de redes complexas. **Transinformação**, Campinas, v. 30, n. 2, 2022. Disponível em: https://periodicos.puc-campinas.edu.br/transinfo/article/view/5963. Acesso em: 15 nov. 2022.

CUNHA, Murilo Bastos da; COSTA, Maira Murrieta. Fontes de informação sobre Gestão de Dados de Pesquisa. **Informação & Sociedade: Estudos**, João Pessoa, v. 30, n. 4, p. 1–59, 2021. DOI: 10.22478/ufpb.1809-4783.2020v30n4.57183. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/57183. Acesso em: 15 nov. 2022.

FALSARELLA, Orandi Mina; JANNUZZI, Celeste Sirotheau Corrêa. Inteligência organizacional e competitiva e big data: uma visão sistêmica para a gestão sustentável das organizações. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 25, n. 1, p. 179–204, 2020. Disponível em: https://periodicos.ufmg.br/index.php/pci/article/view/22658. Acesso em: 15 nov. 2022.

FREUND, Gislaine Parra; FAGUNDES, Priscila Basto; MACEDO, Douglas Dyllon Jeronimo de; DUTRA, Moisés Lima. Mecanismos tecnológicos de segurança da informação no tratamento da veracidade dos dados em ambientes Big Data. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 24, n. 2, p. 124–142, 2019. Disponível em: https://periodicos.ufmg.br/index.php/pci/article/view/22621. Acesso em: 15 nov. 2022.

HENNING, Patricia Corrêa; RIBEIRO, Claudio José Silva; SANTOS, Luiz Olavo Bonino da Silva; SANTOS, Paula Xavier dos. GO FAIR e os princípios FAIR: o que representam para a expansão dos dados de pesquisa no âmbito da Ciência Aberta. **Em Questão**, Porto Alegre, v. 25, n. 2, p. 389–412, 2019. DOI: 10.19132/1808-5245252.389-412. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/84753. Acesso em: 15 nov. 2022.

JESUS, Ananda Fernanda de; CASTRO, Fabiano Ferreira de; RAMALHO, Rogério Apareciso Sá. O papel das bibliotecas no Linked Data. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 26, p. 01–21, 2021. DOI: 10.5007/1518-2924.2021.e75909. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/75909#:~:text=Objetivo%3A%20As%20bibliotecas%20possuem%20um,tecnol%C3%B3gicos%2C%20como%20o%20estabelecimento%20do. Acesso em: 15 nov. 2022.

LEMOS, Robson Rodrigues; GOLÇALVES, Alexandre Leopoldo; SANTOS, César Cardoso; FREITAS, Maykon Carlos. Visualização de dados do exame nacional brasileiro do ensino médio: VisDadosEnem. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 23, n. 53, p. 124–136, 2018. DOI: 10.5007/1518-2924.2018v23n53p124. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2018v23n53p124. Acesso em: 15 nov. 2022.

LOTT, Yuri Monnerat; CIANCONI, Regina de Barros. Vigilância e privacidade, no contexto do Big Data e dados pessoais: análise da produção da Ciência da Informação no Brasil. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 23, n. 4, p. 117–132, 2018. Disponível em: https://periodicos.ufmg.br/index.php/pci/article/view/22594. Acesso em: 15 nov. 2022.

MELLO FILHO, Luiz Lourenço de; ARAÚJO JÚNIOR, Rogério Henrique de. Objetos de fronteira: um diálogo entre a ciência da informação e a ciência de dados. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 26, p. 01–22, 2021. DOI: 10.5007/1518-2924.2021.e77247. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/77247. Acesso em: 15 nov. 2022.

MENDONÇA, Claudio Marcio Campos de; ANDRADE, António Manuel Valente de. Uso da IoT, Big Data e Inteligência Artificial nas capacidades dinâmicas: um estudo comparativo entre cidades do Brasil e de Portugal. Informação & Sociedade: Estudos, João Pessoa, v. 29, n. 4, p. 37–60, 2019. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/47755. Acesso em: 15 nov. 2022.

MESCHINI, Fabio Orsi; FRANCELIN, Marivalde Moacir. Big data e Organização do Conhecimento: reflexões iniciais a partir de uma proposta classificatória da produção científica. **Em Questão**, Porto Alegre, v. 28, n. 1, p. 35–63, 2021. DOI: 10.19132/1808-5245281.35-63. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/113697. Acesso em: 15 nov. 2022.

MONTEIRO, Elizabete Cristina de Souza de Aguiar; SANT'ANA, Ricardo Cesar Gonçalves. Plano de gerenciamento de dados em repositórios de dados de universidades. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 23, n. 53, p. 160–173, 2018. DOI: 10.5007/1518-2924.2018v23n53p160. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2018v23n53p160. Acesso em: 15 nov. 2022.

MONTEIRO, Silvana Drumond. A vida secreta dos metadados no wikidata: um enfoque sobre o sentido na (Web) Semântica Formal. **Informação & Sociedade: Estudos**, João Pessoa, v. 28, n. 1, 2018. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/34757. Acesso em: 15 nov. 2022.

MOREIRA, Fabio Mosso; ALEIXO, Diana Vilas Boas Souto; BISI, Pedro Henrique Santos; FRANCHI, Leonardo Felipe; SANT'ANA, Ricardo César Gonçalves. Construção Colaborativa de Representações para a Disseminação de Dados Agrícolas: Um Estudo do Portal CoDAF . **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 23, n. 52, p. 61–72, 2018. DOI: 10.5007/1518-2924.2017v23n52p61. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2017v23n52p61. Acesso em: 15 nov. 2022.

MORENO, Fernanda Passini. Repositórios de dados de pesquisa na Espanha: breve análise. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 23, n. 53, p. 52–63, 2018. DOI: 10.5007/1518-2924.2018v23n53p52. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2018v23n53p52. Acesso em: 15 nov. 2022.

MOTTA, Fernanda Miranda de Vasconcellos; BARBOSA, Cátia Rodrigues; BARBOSA, R. R. Big Data como fonte de inovação em museus: o estudo de caso do Museu Britânico. **Informação & Sociedade: Estudos**, João Pessoa, v. 29, n. 1, 2019. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/44005. Acesso em: 15 nov. 2022.

NESELLO, Priscila; FACHINELLI, Ana Cristina. The effects of big data over the analytical activities of strategic intelligence professionals in Brazil. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 24, n. 2, p. 87–102, 2019. Disponível em: https://periodicos.ufmg.br/index.php/pci/article/view/22619. Acesso em: 15 nov. 2022.

OLIVEIRA, Caliel Cardoso; SILVA, Maurício Coelho; PAVÃO, Caterina Marta Groposo; SILVA, Fabiano Couto Corrêa; MOURA, Ana Maria Mielniczuk; BARROS, Thiago Henrique Bragato. A teoria da citação de dados: uma revisão da produção científica na América Latina. **Transinformação**, Campinas, v. 34, p. 1–18, 2022. Disponível em: https://periodicos.puc-campinas.edu.br/transinfo/article/view/6487. Acesso em: 15 nov. 2022.

OUCHI, Marcos Teruo; ARAKAKI, Ana Carolina Simionato. Um estudo da Blockchain aplicado ao contexto dos Dados de Pesquisa. **Em Questão**, Porto Alegre, v. 26, n. 3, p. 70–93, 2020. DOI: 10.19132/1808-5245263.70-93. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/98345. Acesso em: 15 nov. 2022.

PEREIRA, Clayton Martins; FERNEDA, Edberto; SANTAREM SEGUNDO, José Eduardo. Análise do processo de recuperação da informação em bases de dados publicadas como dados abertos ligados utilizando a abordagem RDB2LOD. **Em Questão**, Porto Alegre, v. 26, n. 3, p. 94–120, 2020. DOI: 10.19132/1808-5245263.94-120. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/98764. Acesso em: 15 nov. 2022.

PINTO, Vitor Afonso; CARDOSO, Ana Maria Pereira; PINHEIRO, Marta Macedo Kerr; PARREIRAS, Fernando Silva. Iinterdisciplinarity in Data Science over Big Data: findings for mining industry. **Informação & Sociedade: Estudos**, João Pessoa, v. 29, n. 4, p. 61–74, 2019. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/47536. Acesso em: 15 nov. 2022.

RAUTENBERG, Sandro; BURDA, Alessandra Cassiana; SOUZA, Lucélia de. Um workflow para compartilhamento de dados científicos primários baseado em dados abertos conectados. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 23, n. 53, p. 110–123, 2018. DOI: 10.5007/1518-2924.2018v23n53p110. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2018v23n53p110. Acesso em: 15 nov. 2022.

ROCHA, Rafael Port da; CAREGNATO, Sônia Elisa; GABRIEL JUNIOR, Rene Faustino. Aspectos de inovação na implantação de um centro de digitalização e gestão de dados da pesquisa. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, Edição Especial - v. 23, n. esp. 1, p. 1–15, 2018. DOI: 10.5007/1518-2924.2018v23nespp1. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2018v23nespp1. Acesso em: 15 nov. 2022.

SANTOS-D'AMORIM, Karen Isabelle dos; CRUZ, Rúbia Wanessa dos Reis; SILVA, Marcela Lino da; CORREIA, Anna Elizabeth Galvão Coutinho. Dos dados ao conhecimento: tendências da produção científica sobre Big Data na Ciência da Informação no Brasil. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 25, p. 01–23, 2020. DOI: 10.5007/1518-2924.2020.e70527. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2020.e70527. Acesso em: 15 nov. 2022.
SILVA, Sivaldo Pereira da; SOARES, Ana Thereza Nogueira; CESAR, Daniel Jorge Teixeira; RABELO, Leon Eugênio Monteiro. Indicadores para avaliação qualitativa de Dados Abertos: Inteligibilidade, operacionalidade e interatividade nos datasets do Governo Federal no Portal Brasileiro de Dados Abertos. **Informação & Sociedade: Estudos**, João Pessoa, v. 30, n. 3, p. 1–19, 2020. DOI: 10.22478/ufpb.1809-4783.2020v30n3.52469. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/52469. Acesso em: 15 nov. 2022.
SILVA, Walison Dias da; PARREIRAS, Fernando Silva; MAIA, Luiz Cláudio Gomes; BRANDÃO, Wladmir Cardoso. Anotação semântica automática do currículo Lattes utilizando Linked Open Data. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 23, n. 4, p. 53–72, 2018. Disponível em: https://periodicos.ufmg.br/index.php/pci/article/view/22591. Acesso em: 15 nov. 2022.
SILVEIRA, Lúcia da; BARBOSA, Amanda Dall'Agnol; FERREIRA, Manuela Klanovicz; CAREGNATO, Sônia Elisa. Citação de dados científicos: scoping review. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 25, p. 01–31, 2020. DOI: 10.5007/1518-2924.2020.e72153. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2020.e72153. Acesso em: 15 nov. 2022.
SIMIONATO, Ana Carolina; CONEGLIAN, Caio Saraiva; GONÇALEZ, Paula Regina Ventura Amorim; SANTARÉM SEGUNDO, José Eduardo. Audiovisuais e Linked data: um estudo das bases DBpedia e LMDB. **Em Questão**, Porto Alegre, v. 24, n. 3, p. 297–315, 2018. DOI: 10.19132/1808-5245243.297-315. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/78206. Acesso em: 15 nov. 2022.
SOBRAL, Natanael Vitor; LIMA, Gillian Leandro de Queiroga; SOBRAL, Ana Sara Pereira de Melo. Produção científica sobre hospitais no contexto da ciência de dados: um estudo a partir da web of science. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 26, n. Especial, p. 1–16, 2021. DOI: 10.5007/1518-2924.2021.78824. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/78824. Acesso em: 15 nov. 2022.
SZINVELSKI, Mártin Marks; ARCENO, Taynara Silva; FRANCISCO, Lucas Baratieri. Perspectivas jurídicas da relação entre big data e proteção de dados. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 24, n. 4, p. 132–144, 2019. Disponível em: https://periodicos.ufmg.br/index.php/pci/article/view/22644. Acesso em: 15 nov. 2022.
TORINO, Emanuelle; TREVISAN, Gustavo Lunardelli; CONEGLIAN, Caio Saraiva; BOTEGA, Leonardo Castro; SANTARÉM SEGUNDO, José Eduardo; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Enriquecimento semântico para a disponibilização de dados abertos: teoria e prática. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 25, p. 01–19, 2020. DOI: 10.5007/1518-2924.2020.e67991. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2020.e67991. Acesso em: 15 nov. 2022.
TORINO, Emanuelle; VIDOTTI, Silvana Aparecida Borsetti Gregorio; VECHIATO, Fernando Luiz. Contribuições do atributo Metadados para a Encontrabilidade da Informação. **Em Questão**, Porto Alegre, v. 26, n. 2, p. 437–457, 2020. DOI: 10.19132/1808-5245262.437-457. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/93072. Acesso em: 15 nov. 2022.
VICTORINO, Marcio de Carvalho; MARTINS, Luiz; HOLANDA, Maristela; FONSECA, Rodrigo. Arquitetura de publicação de dados abertos conectados governamentais da Universidade de Brasília. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 25, p. 01–25, 2020. DOI: 10.5007/1518-2924.2020.e67665. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2020.e67665. Acesso em: 15 nov. 2022.

Fonte: Elaborado pelos autores (2022)

Apêndice D – Nuvem de palavras

Fonte: Elaborado pelos autores (2022)

CLASSIFICAÇÃO AUTOMÁTICA DE ARTIGOS PUBLICADOS NO ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO - 2021

Automatic Classification of Articles Published in the National Meeting on Research in Information Science - 2021

Liliane Cristina Soares Sousa[1], José Eduardo Santarém Segundo [(2),] Fábio Parra Furlanete[(3)]
(1) Universidade Estadual de Londrina - UEL, Londrina/PR, lilianeli.sousa@uel.br
(2) Universidade de São Paulo – USP, Ribeirão Preto/SP, santarem@usp.br
(3) Universidade Estadual de Londrina – UEL, Londrina/PR, ffurlanete@uel.br

Resumo
A classificação de documentos científicos, diante do número expressivo de produção e disponibilização informacional, é um desafio contemporâneo para a Ciência da Informação. Diante disso, propõe-se fazer um experimento com algoritmos de Machine Learning, para entender de que maneira essas ferramentas, podem processar dados textuais e gerar métricas de validação. A acurácia, é uma métrica de validação para entender este processo de classificação na análise de dados não estruturados, como os documentos textuais. A metodologia se baseia nas concepções de Mineração de Textos, que dispõe de ferramentas para estruturar análise de dados não estruturados ou semiestruturados. Ao aplicar diferentes algoritmos de Machine Learning na massa de textos, requer observar o comportamento destes no processamento dos dados. A pesquisa preliminar foi promissora, no que se refere, a entender como se comporta a aplicação de algoritmos de Machine Learning em dados textuais. Conclui-se que, estudos em torno da área de Data Science, particularmente usando algoritmos de Machine Learning podem ser aprofundados e aplicados na Ciência da Informação, com o propósito de agregar valor para diversas áreas da ciência, em particular, para os profissionais da informação.

Palavras-chave: Ciência da Informação; Machine Learning; Classificação textual; Dados não-estruturados; Algoritmo.

Abstract: The classification of scientific documents, given the significant number of production and information availability, is a contemporary challenge for Information Science. Therefore, it is proposed to do an experiment with Machine Learning algorithms, to understand how these tools can process textual data and generate validation metrics. Accuracy is a validation metric to understand this classification process in the analysis of unstructured data, such as textual documents. The methodology is based on the concepts of Text Mining, which has tools to structure analysis of unstructured or semi-structured data. When applying different Machine Learning algorithms in the mass of texts, it is necessary to observe their behavior in the data processing. Preliminary research was promising, in terms of understanding how the application of Machine Learning algorithms in textual data behaves. It is concluded that studies around the area of Data Science, particularly using Machine Learning algorithms, can be deepened and applied in Information Science, with the purpose of adding value to several areas of science, in particular, for information professionals.

Keywords: Information Science; Machine Learning; Text classification; Unstructured data; Algorithm.

1. Introdução

O presente texto faz parte de uma investigação preliminar, para compreender se algoritmos de Machine Learning são capazes de obter uma boa acurácia (CA) no processo de classificação textual. Importante destacar que Machine Learning, é o aprendizado de máquina, que estabelece ferramentas de análise de dados, que permite a automação de elaboração de modelos analíticos de processamento de dados. Como observa Jordan e Mitchell (2015, p. 255, tradução nossa), “O estudo do aprendizado de máquina é importante tanto para abordar questões científicas e de engenharia fundamentais, quanto para software de computador altamente práticos que são produzidos e utilizados em diversos aplicativos.” Diante disso, deseja-se testar a importância de estudos de Data Science e sua aplicação na Ciência da Informação, para auxiliar no processo de análise de

corpus textual. Compreende-se que os estudos de Data Sciense que envolvem as informações, o processo de seleção, preparação, transformação, desenvolvimento, processamento e análise de dados, são significativos para a Ciência da Informação.

O corpus selecionado está para uma perspectiva de análise de dados não-estruturados. Sendo que, o objeto da investigação está direcionado aos trabalhos publicados no ENANCIB 2021.

O Encontro de Pesquisa em Ciência da Informação (ENANCIB), representa o principal evento de pesquisa e de pós-graduação da área de Ciência da Informação do país. Aspira problematizar e refletir a produção científica da área, visto que, requer fomentar amplo diálogo entre os pesquisadores protagonistas nos programas de pós-graduação. O evento é direcionado para estimular troca de experiências acadêmico-científicas e pela solidificação de laços acadêmicos no âmbito nacional e internacional.

A presente investigação trará a aplicação de técnicas de Mineração de Texto, para alcançar a hipótese projetada. A Mineração de Texto, segundo Barion e Lago (2008, p. 125), é identificada como uma extensão da área de Data Mining; e de acordo com Thuraisinggham (1999, p. 210), tem por objetivo central extrair modelos e associações singulares de uma gama significativa de um banco de dados textual. Ainda na concepção de Barion e Lago (2008, p. 123), esses processos possibilitam usar conjuntos de estratégias para navegar, organizar, achar e descobrir informações em base de textos.

No dizer de Wives (2004, p. 66) é enfatizado que a maneira básica de Mineração de Texto, está relacionado na exploração e identificação de terminologias relevantes na composição de um grupo de documentos, como também instituir padrões textuais e projetar grupos temáticos de conteúdos pela periodicidade de aparecimento de termos no corpus a ser investigado. Esse processo de mineração de texto, se caracteriza como um método de suporte para pesquisadores, a fim de possibilitar um novo prisma informacional, em grande e significativa coleção de textos.

A Mineração de Texto emerge em consequência da necessidade de explorar, de maneira automática, modelos e anomalias informacionais em documentos. Segundo Aranha e Passos (2006, p. 125), a Mineração de Texto, é uma área do conhecimento multidisciplinar, que permeia diversos campos da ciência, como “Informática, Estatística, Linguística e Ciência Cognitiva”, conforme esta investigação preliminar, entende-se que a Ciência da Informação, também é um campo científico de atuação na Mineração de Texto.

Nossa questão é: Em que medida algoritmos de Machine Learning são capazes de processar métricas de validação em classificação de dados textuais?

2. Objetivos

Este artigo tem o objetivo de diagnosticar se algoritmos de Machine Learning, aplicados em dados não-estruturados, são capazes de classificar documentos textuais, pelo olhar das técnicas de classificação dos bibliotecários. Tendo como corpus de pesquisa os artigos publicados no ENANCIB 2021.

3. Procedimentos Metodológicos

A execução dessa investigação fundamenta-se na Mineração de Texto, uma vez que, estabelece procedimentos de extração de padrões em significativas quantidades de textos, tendo como atributo a linguagem natural, e geralmente, para ser utilizados com objetivo específico. Visto que, a Mineração de Texto, requer selecionar conhecimentos relevantes de dados não estruturados ou semiestruturados.

As etapas desenvolvidas nesta pesquisa, estão representadas a seguir.

O primeiro passo para atingirmos o objetivo proposto, foi compreender os processos necessários para a aplicação da Mineração de Texto, em particular, nos dados selecionados para esta investigação. Seguindo a concepção de Aranha (2007, p. 150), a Mineração de Texto, está permeada por quatro macro etapas: “coleta, pré-processamento, indexação e análise da informação”.

Figura 1 – Etapas do processo de análise dos dados

Fonte: Os autores (2022)

A posteriori, iniciamos a coleta dos dados, nos Anais do ENANCIB - 2021. Essa etapa, segundo Schiessl e Medeiros (2011, p. 100), consiste na elaboração do objeto da pesquisa, formada na Mineração de Texto, por uma base textual, no qual será trabalhada no processo de análise. Pode-se denominar esta base de textos, como Corpus; composto por um grupo de textos, que representa um conjunto de linguagens naturais. O corpus desta investigação foi extraído do Anais do ENANCIB - 2021, foram selecionados 342 artigos, separados por GTs: GT1 (19); GT2 (38); GT3 (38); GT4 (49); GT5 (42); GT6 (32); GT7 (26); GT8 (43); GT9 (15); GT10 (25); GT11 (15).

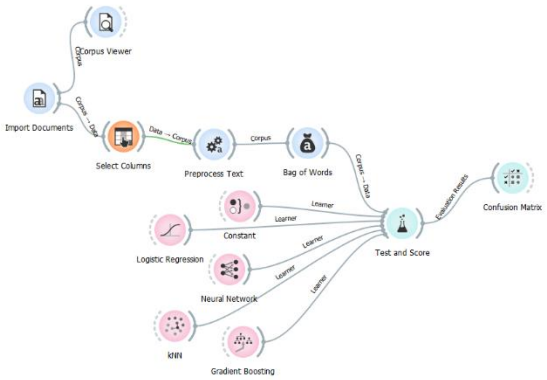
Posteriormente, na segunda etapa denominada pré-processamento, que é caracterizada por ser responsável pela construção de uma estrutura representativa dos documentos textuais. No qual pretende-se com este processamento, aprimorar a organização e a qualidade dos dados. Esta ação consiste em aplicar diversas maneiras de transformações nos documentos textuais, com a ideia de estruturar estes grupos de documentos, para poderem ser submetidos a algoritmos de mineração de texto. No dizer de Aranha e Passos (2006, 132), as estratégias utilizadas nos "tratamentos dados ao texto durante essa fase podem ser feitos tanto de forma automatizada como feitos por humanos, porém o desempenho dos sistemas automáticos é extremamente superior." E segundo Spark-Jones e Willet (1997, p. 5), essa etapa de pré-processamento "inclui tokenização, limpeza de dados e eliminação de stopwords." Nesta fase, foi analisado a necessidade de preparar os textos para que a observação seja eficiente. Identificou-se que a massa documental estava desbalanceada, percebendo que em cada delimitação dos grupos de trabalho, há uma quantidade diferente de artigos. Deste modo, opta-se por equilibrar a massa textual, para obter um mapeamento equitativo dos dados, sem enviesar os resultados pelo desbalanceamento das categorias. Isto posto, os dados textuais foram reorganizados, equiparando cada grupo de trabalho do ENANCIB 2021, totalizando 165 artigos, distribuindo 15 artigos por GT (GT1 / GT2 / GT3 / GT4 / GT5 / GT6 / GT7 / GT8 / GT9 / GT10 / GT11).

Ainda nesta perspectiva de pré-processamento dos dados, como terceira etapa do processo, tem-se a indexação dos dados, que objetiva uma acessibilidade rápida e eficiente destas informações, a busca por palavras. Faz parte da Mineração dos dados, onde se seleciona a tarefa que será executada, conforme a necessidade desejada.

O próximo passo refere-se à construção de uma estrutura de dados, composta pelas etapas anteriores, no qual, aplica-se algoritmos de mineração de dados, com o objetivo de extrair os conhecimentos. Por fim, é feita a análise da aplicação dos algoritmos, entrando na etapa de análise e leitura dos dados gerados.

Todos os processos embasados teoricamente, expostos anteriormente, foram seguidos para obter-se os resultados preliminares de análise. As etapas seguidas para a execução do mapeamento do processamento dos textos, estão ilustradas na Figura 2:

Figura 2 – Mapeamento do processo de execução da Mineração de Texto



Fonte: Elaborado pelos autores usando o software Orange Canvas (2022)

4. Resultados

Nesse passo, processa-se os dados textuais com objetivo de identificar a acurácia de acerto de classificação pelos algoritmos. O escopo da acurácia para o processamento e leitura dos dados, fornece-nos uma junção entre exatidão e precisão, consequentemente, oferece-nos resultados próximos do valor real das classificações. O fator selecionado para o processamento dos dados foi a validação cruzada, com número dez de dobras e 70% para treino com 30% dos dados para teste.

Para a construção das amostragens, foram utilizados os seguintes algoritmos: KNN, executa previsões conforme as classes mais próximas, utilizou-se cinco como vizinhos mais próximos, sendo a métrica euclidiana e peso uniforme; Redes Neurais (Scikit-Learn), usando Perceptron multicamadas, que é capaz de aprender modelos não lineares e lineares, sendo parâmetro utilizado para esse algoritmo 200 como número máximo de interações; Logistic Regression, essa ferramenta representa o algoritmo de classificação de regressão logística com regularização, onde utilizou-se o tipo de regularização Ridge (L2); Gradient Boosting, algoritmo que utiliza a combinação de resultados de preditores fracos, com a intenção de elaborar modelos preditivos, os parâmetros aplicados foram de 100 para o número de árvores e taxa de aprendizagem de 0,100 e Constant, que gera uma amostra o qual na maioria das vezes prevê a maior parte para tarefas de classificação e valor médio para funções de regressão.

Posterior a execução dos testes, com os diversos algoritmos de Machine Learning, foi selecionado o algoritmo Logistic Regression. A seleção desse algoritmo, fundamenta-se, pela sua característica de processamento e visualização das relações entre os dados. Em face do exposto, elegeu-se a acurácia como métrica de validação para a presente observação. Visto que, a CA (acurácia) significa a proximidade de resultado levando em consideração o seu valor de referência real, ou seja, quanto maior o nível de acuracidade, mais perto do modelo significa o resultado encontrado.

Dessa maneira, segue-se com os resultados dos testes aplicados nos dados textuais selecionados e pré-processados. Observa-se a figura 3.

Figura 3 – Análise do algoritmo selecionado

Model	AUC	CA	F1	Precision	Recall
kNN	0.932	0.952	0.733	0.733	0.733
Neural Network	0.970	0.945	0.710	0.688	0.733
Logistic Regression	0.960	0.952	0.714	0.769	0.667
Gradient Boosting	0.927	0.939	0.643	0.692	0.600
Constant	0.500	0.909	0.000	0.000	0.000

Fonte: Elaborado pelos autores usando o software Orange Canvas (2022)

Ao extrair os dados processados no teste com o algoritmo Logistic Regression, verifica-se que a CA (acurácia) extraída foi de 95%.

Na figura 4, é possível fazer a leitura que o algoritmo Logistic Regression representou na Mineração dos textos selecionados.

Para executar a leitura destes dados gerados, utilizou-se a Confusion Matrix, que apresenta o número ou proporção entre a classe prevista e a atual. Esses resultados foram consequência dos dados gerados pela avaliação do algoritmo Logistic Regression.

Figura 4 – Análise do processamento dos dados

	GT8	GT10	GT4	GT1	GT5	GT3	GT11	GT2	GT9	GT7	GT6	Σ
GT8	7	0	0	1	0	0	1	1	1	3	1	15
GT10	0	10	0	0	1	1	0	0	2	1	0	15
GT4	1	0	11	0	1	0	0	0	0	0	2	15
GT1	1	0	0	10	1	1	0	0	0	0	2	15
GT5	1	1	0	1	9	1	1	0	0	1	0	15
GT3	2	0	3	0	0	7	1	0	0	0	2	15
GT11	1	0	1	0	1	0	9	0	0	2	1	15
GT2	2	0	0	2	0	0	0	11	0	0	0	15
GT9	2	2	0	0	0	0	0	0	11	0	0	15
GT7	0	0	0	0	0	0	1	0	0	13	1	15
GT6	0	0	0	2	0	1	0	0	0	1	11	15
Σ	17	13	15	16	13	11	13	12	14	21	20	165

Fonte: Elaborado pelos autores usando o software Orange Canvas (2022)

Visualiza-se na figura 4, o processamento dos dados, e indicam a seguinte leitura:

- GT1 (15 artigos), 10 artigos estão conforme respectiva classe, no entanto, existem: 1 artigo para o GT8, 1 artigo para o GT5, 1 artigo para o GT3 e 2 artigos para o GT6. Dessa maneira, de acordo com a previsão de proporção, 66,66% dos

artigos do GT1 pertencem a sua classificação prévia;

- GT2 (15 artigos), 11 artigos estão conforme respectiva classe, no entanto, existem: 2 artigos para o GT8 e 2 artigos para o GT1. Sendo assim, de acordo com a previsão de proporção, 73,33% dos artigos do GT2 pertencem a sua classificação prévia;
- GT3 (15 artigos), 7 artigos estão conforme respectiva classe, no entanto, existem: 2 artigos para o GT8, 3 artigos para o GT4, 1 artigo para o GT11 e 2 artigos para o GT6. Dessa maneira, de acordo com a previsão de proporção, 46,66% dos artigos do GT3 pertencem a sua classificação prévia;
- GT4 (15 artigos), 11 artigos estão conforme a respectiva classe, no entanto, existem: 1 artigo para o GT8, 1 artigo para o GT5 e 2 artigos para o GT6. Dessa maneira, de acordo com a previsão de proporção, 73,33% dos artigos do GT4 pertencem a sua classificação prévia;
- GT5 (15 artigos), 9 artigos estão conforme a respectiva classe, no entanto, existem: 1 artigo para GT8, 1 artigo para GT10, 1 artigo para GT1, 1 artigo para GT3, 1 artigo para o GT11 e 1 artigo para GT7. Dessa maneira, de acordo com a previsão de proporção, 60% dos artigos do GT5 pertencem a sua classificação prévia;
- GT6 (15 artigos), 11 artigos estão conforme respectiva classe, no entanto, existem: 2 artigos para o GT1, 1 artigo para o GT3 e 1 artigo para o GT7. Sendo assim, de acordo com a previsão de proporção, 73,33% dos artigos do GT6 pertencem a sua classificação prévia.
- GT7 (15 artigos), 13 artigos estão conforme a respectiva classe, no entanto, existem: 1 artigo para o GT11 e 1 artigo para o GT6. Sendo assim, de acordo com a previsão de proporção, 86,66% dos artigos do GT7 pertencem a sua classificação prévia;
- GT8 (15 artigos), 7 artigos estão conforme a respectiva classe, no entanto, existem: 1 artigo para o GT1, 1 artigo para o GT11, 1 artigo para o GT2, 1 artigo para o GT9, 3 artigos para o GT7 e 1 artigo para o GT6. Dessa maneira, de acordo com a previsão de proporção, 46,66% dos artigos do GT8 pertencem a sua classificação prévia;
- GT9 (15 artigos), 11 artigos estão de acordo com a respectiva classe, no entanto, existem: 2 artigos para o GT8 e 2 artigos para GT10. Sendo assim, de acordo com a previsão de proporção, 73,33% dos artigos do GT9 pertencem a sua classificação prévia;
- GT10 (15 artigos), 10 artigos estão conforme respectiva classe, no entanto, existem: 1 artigo para o GT5, 1 artigo para o GT3, 2 artigos para o GT9 e 1 artigo para o GT7. Dessa forma, de acordo com a previsão de proporção, 66,66% dos artigos do GT10 pertencem a sua classificação prévia;
- GT11 (15 artigos), 9 artigos estão conforme a respectiva classe, no entanto, existem: 1 artigo para o GT8, 1 artigo para o GT4, 1 artigo para o GT5, 2 artigos para o GT7 e 1 artigo para o GT6. Sendo assim, de acordo com a previsão de proporção, 60% dos artigos do GT11 pertencem a sua classificação prévia;

O processamento dos dados com o algoritmo Machine Learning, demonstrou a possibilidade de validação, no processamento de mineração de texto com os cinco algoritmos testados, inclusive, com o Logistic Regression, selecionado para a apresentação dos índices de CA da análise.

A mineração de textos indicou, nesta perspectiva, uma estratégia importante, cuja intenção, de acordo com Feldman e Hirsh, citados por Wives (2004, p. 80), significa "constituir-se em um meio efetivo de recuperação, filtragem, manipulação e resumo do conhecimento contido em grandes volumes de informações textuais, para apresentá-lo em forma de gráficos, listas ou tabelas para o consumo de suas informações."

5. Considerações Finais

Esta pesquisa preliminar, identificou que analisar informações em documentos não estruturados significa um desafio. Mesmo com os avanços no âmbito da tecnologia informacional, salienta-se a necessidade de trilhar caminhos de preparação dos documentos textuais, a fim de qualificar os dados, para obter resultados no processamento desses dados.

Diante dos resultados obtidos, a partir do processamento dos dados textuais,

demonstrou-se a viabilidade de aprofundamento na investigação. A possibilidade de entender quais os elementos foram levados em consideração na classificação dos textos feitas pelo algoritmo.

É interessante observar que ao aumentar o volume de dados a ser testado, pode-se conseguir identificar elementos interessantes de análise, como a GTs de trabalho que podem ter conteúdos mais aderentes entre eles, ou ainda grupos de trabalho que tenham termos mais significativos que os colocam em situação de exclusividade de tema em relação a outros GTs.

A mineração de textos, a partir das concepções de Data Science, permite a construção de ferramentas de extração e uso de dados, que têm a fortalecer as pesquisas na Ciência da Informação.

Referências

ARANHA, Christian; PASSOS, Emmanuel. A Tecnologia de Mineração de Textos. **Revista Eletrônica de Sistemas de Informação**, [S.l.], v. 5, n. 2, ago. 2006. ISSN 1677-3071. Disponível em: http://periodicosibepes.org.br/index.php/reinfo/article/view/171>. Acesso em: 12 set. 2022. doi:https://doi.org/10.21529/RESI.2006.0502001.

ARANHA, Christian N. **Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português**: Sob o Enfoque da Inteligência Computacional. 2007. 144f. Tese (Doutorado) - Programa de Pós-graduação em Engenharia Elétrica, Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, 2007. Disponível em: https://www.maxwell.vrac.puc-rio.br/10081/10081_4.PDF.Acesso em 03 jun. 2022.

BARION, E. C.N. **Mineração de textos**. 2008. Disponível em: https://revista.pgsskroton.com.br/index.php/rcext/article/view/2372/2276. Acesso em: 24 nov. 2018.

INGERSOLL, Grant S.; MORTON, Thomas S.; FARRIS, Andrew L. 2013. **Taming Text: How to find, organize and manipulate it.** Shelter Island, NY (USA): Manning Publications Co., 2013. 298p.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255-260, 2015. Disponível em: <https://cs.uwaterloo.ca/~y328yu/mycourses/480-2018/readings/JordanMitchell.pdf>. . Acesso em: 20 nov. 2022.

SPARK-JONES, K; WILLET, P. (1997). **Readings in Information Retrieval**. Morgan Kaufmann. 1997.

SCHIESSL, M.; MEDEIROS, M. Descoberta de Conhecimento em Texto aplicada a um Sistema de Atendimento ao Consumidor. **Revista Ibero-Americana de Ciência da Informação**, v. 4, n. 2, p. 94-111, 2011.

THURAISINGHAM, Bhavani M. Data mining: technologies, techniques, tools, and trends. Boca Raton: Editora CRC Press, 1999. 288p.

WIVES, L. K. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos**. Tese (Doutorado em Computação) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2004.

CONTRATOS INTELIGENTES NO DESENVOLVIMENTO DE COLEÇÕES: UMA ABORDAGEM ORIENTADA À *BLOCKCHAIN*

SMART CONTRACTS IN COLLECTION DEVELOPMENT: A BLOCKCHAIN-ORIENTED APPROACH

Rafael Rocha[1], Gercina Ângela de Lima[2]
(1) Programa de Pós-Graduação em Gestão & Organização do Conhecimento - UFMG, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, rafael-rocha@ufmg.br
(2) Programa de Pós-Graduação em Gestão & Organização do Conhecimento - UFMG, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, limagercina@gmail.com

Resumo
O desenvolvimento de coleções é um processo cíclico, cujo objetivo é manter uma biblioteca alinhada aos anseios informacionais de sua comunidade. Para a garantia desse processo, diversos métodos foram propostos, mas nenhum deles adotou uma abordagem pragmática que envolvesse os contratos inteligentes na *blockchain*. Nesse contexto, esta pesquisa tem como objetivo buscar um método orientado à *blockchain*, com uso dos contratos inteligentes nas etapas do desenvolvimento de coleções para auxiliar os profissionais da informação no gerenciamento de um acervo bibliográfico. A metodologia se caracteriza como exploratória e descritiva, a partir da literatura disponível na área sobre os recursos oferecidos pela *blockchain*, e se desenvolve em quatro etapas: busca e seleção dos documentos, análise e seleção de repositórios públicos de código, implementação dos contratos inteligentes em uma *blockchain* criada para o desenvolvimento de coleções e avaliação da solução proposta. Como resultados parciais, chegou-se à identificação dos principais recursos necessários em cada etapa do processo e das diversas implementações de contratos inteligentes. Por fim, a pesquisa em andamento revelou-se promissora, robusta e adequada para ser desenvolvida.

Palavras-chave: biblioteca; desenvolvimento de coleções; política de desenvolvimento de coleções; *blockchain*; contratos inteligentes.

Abstract
The development of collections is a cyclical process, whose objective is to keep a library updated to the informational desires of its community. To ensure this process, several proposed methods were proposed, but none of their methods was a pragmatic approach involving smart contracts on the blockchain. In this context, this research aims to seek a blockchain-oriented method, using smart contracts in the stages of collection development to assist information professionals in managing a bibliographic collection. The methodology is characterized by the characterization of the code area, from the selection available on the selection code, from the selection of resources offered by the selection, and by the selection of resources available for the selection of resources developed by the selection and by the selection of the selection documents developed for the implementation of contracts a blockchain created for the development of collections and evaluation of the proposed solution. As partials, the main features and results were identified at each stage of the process of the various implementations of smart contracts. Finally, the research in progress has proved to be promising, robust and suitable for further development.

Keywords: library; collection development; collection development policy; blockchain; smart contracts.

1 Introdução

O desenvolvimento de coleções é uma atividade realizada pelos bibliotecários para atender às necessidades informacionais dos usuários (SANJAY, 2016). As coleções são oriundas de materiais físicos ou digitais, podendo ser desde uma gravação de áudio até um livro. Conforme Reitz (2004), é necessário utilizar a análise estatística e as projeções demográficas, considerando as restrições orçamentárias. Ademais, o desenvolvimento de coleções engloba diversas etapas, tais como: seleção, aquisição, avaliação, debastamento, dentre outras. Cada etapa é regida pela Política de Desenvolvimento de Coleções (PDC) da instituição, cujo foco deve ser as necessidades do usuário, pois é ele quem utiliza os recursos informacionais; portanto, o usuário está em posição preferível para julgar a coleção da biblioteca.

Conforme Vignau e Meneses (2005), o desenvolvimento de coleções é o processo mais significativo e efetivo para qualquer biblioteca. O intuito desse processo é manter o crescimento das coleções de modo sustentável, uma vez que há acúmulo exponencial de informação (VERGUEIRO, 1993). A ação de gerir as coleções caracteriza-se como um processo cíclico que transpassa a mera acumulação de material, sendo orientado por uma política que assegura o planejamento e a tomada de decisão (WEITZEL, 2012).

Para tanto, a PDC estabelece as diretrizes que norteiam as ações e os critérios da escolha do acervo. Além disso, atua como contrato entre a comunidade e a biblioteca (JOHNSON, 2004). Essa transparência fortalece o vínculo entre a unidade de informação e o usuário, além de aumentar a participação da comunidade, uma vez que a PDC define a lógica sobre a qual as coleções serão selecionadas (WEITZEL, 2012). Conforme Evans e Saponaro (2005), diversos critérios e objetivos podem ser determinados, tais como: idioma, data de publicação, assunto, frequência de citação, índices, entre outros. Esses critérios contribuem para auxiliar o bibliotecário a formalizar, de modo explícito, as decisões nas diversas etapas do desenvolvimento de coleções.

A evolução das tecnologias possibilitou novas perspectivas para o desenvolvimento de coleções. Segundo Johnson (2004, p. 39), "o futuro do desenvolvimento de coleções será influenciado pela rápida disseminação da tecnologia digital como meio de criação, acesso e entrega de informações", sendo a *blockchain* uma inovação para a área.

Nakamoto (2008) desenvolveu um sistema financeiro baseado em blocos com transações validadas e com marcação horária (*timestamp*). Esse novo modelo de gerir transações ficou conhecido como *blockchain*. O contrato inteligente (*smart contract*) é uma das funcionalidades disponíveis em uma *blockchain* (ALHARBY; VAN MOORSEL, 2017). Os contratos inteligentes são trechos de código registrados na *blockchain*, de natureza pública e auditável. Isso permite que as partes interessadas saibam, previamente, como ocorrerá determinada transação.

No contexto das bibliotecas, Hoy (2017) propôs a utilização da *blockchain* para a gestão de direitos autorais digitais. Essa proposta permite que os materiais digitais sejam identificados unicamente, controlados e transferidos. Assim, o detentor dos direitos autorais dos documentos consegue controlar sua movimentação. Verma (2021) ampliou a proposta, que passou a ser utilizada no compartilhamento de objetos, ferramentas e serviços. Com essa estratégia, segundo esse autor, a biblioteca pode criar moedas e *vouchers* para o financiamento, permitindo manter os registros autenticados e com a garantia de proveniência.

De acordo com Safdar *et al*. (2022), majoritariamente, a literatura analisa, de modo abrangente, as soluções da *blockchain* para a Biblioteconomia. Entende-se que é necessário um enfoque mais pragmático nas soluções da área. Desse modo, o problema é que o desenvolvimento de coleções não adota plenamente novas tecnologias, mais especificamente, os contratos inteligentes. Nesse contexto, questiona-se: quais as contribuições dos contratos inteligentes para o desenvolvimento de coleções?

2 Objetivos

O objetivo principal deste trabalho é identificar as principais contribuições da *blockchain* para o desenvolvimento de coleções. Além disso, almeja-se explorar, do ponto de vista pragmático, as tecnologias em torno dos contratos inteligentes. Os objetivos específicos são: mapear as funcionalidades salientes para o desenvolvimento de coleções no contexto da *blockchain*; identificar, em repositórios públicos de códigos, recursos para a utilização de uma *blockchain* e para a criação de contratos inteligentes; elaborar diversos contratos inteligentes para cada etapa e aplicáveis no desenvolvimento de coleções; e avaliar o uso dos contratos inteligentes na perspectiva do desenvolvimento de coleções.

Por fim, a pesquisa em andamento pretende construir uma *blockchain* com os recursos salientes relatados na literatura científica. Para essa construção, também serão considerados o reúso dos projetos de *blockchain* e os contratos inteligentes

hospedados em repositórios públicos, como no *Github*[1].

3 Procedimentos metodológicos

Seguindo a orientação de Gil (2002), esta pesquisa se caracteriza como exploratória e descritiva, no que diz respeito ao objetivo geral. Isso porque se recorre à literatura científica para obter as informações necessárias sobre a utilização dos contratos inteligentes no desenvolvimento de coleções.

O desenho inicial dos procedimentos metodológicos é dividido em quatro etapas. Na primeira etapa, são coletadas, por meio da literatura científica, as contribuições da *blockchain* e dos contratos inteligentes no desenvolvimento de coleções. A revisão da literatura é conduzida mediante a consulta ao Portal de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)[2] por meio do acesso remoto via Comunidade Acadêmica Federada (CAFe)[3]. Esse Portal concentra a publicação de diversos periódicos. Na busca, foram utilizados os termos: *blockchain, smart contract,* contratos inteligentes, *collection development*, *collection development policy,* desenvolvimentos de coleções, política de desenvolvimento de coleções, *library* e biblioteca. Ademais, operadores lógicos foram utilizados para filtrar os resultados.

Na segunda etapa, é realizada a análise dos repositórios públicos de código, tais como: o Github e o Gitlab[4]. Em ambos, foram buscados os termos: *blockchain* e *smart contract*. Os repositórios devem ter as melhores interações, a saber: quantidade de favoritos, quantidade de *fork*[5] e quantidade de *commits*[6]. Os projetos selecionados servirão de *benchmark* na criação da *blockchain* e na construção dos contratos inteligentes.

A terceira e a quarta etapas ainda não foram aplicadas, pois dependem dos resultados das etapas anteriores. Assim, pretende-se implementar os contratos inteligentes em uma *blockchain* criada para o desenvolvimento de coleções (etapa 3) e realizar a avaliação do conjunto da solução proposta (etapa 4).

4 Resultados

O protocolo da revisão está em estruturação. No entanto, resultados preliminares foram alcançados com a aplicação das duas primeiras etapas. Foram recuperados estudos correlacionando as contribuições dos contratos inteligentes para a Biblioteconomia e Ciência da Informação, mais especificamente, para o campo do desenvolvimento de coleções.

4.1 Desenvolvimento de coleções

Conforme Sanjay (2016), a coleção de uma biblioteca origina-se a partir do conjunto de materiais em diferentes formatos, a saber: mídia imprensa (livros, periódicos e publicações institucionais ou governamentais), microformatos, gravações de áudio e vídeo, dentre outros.

Na literatura, existem algumas sugestões de procedimento para o desenvolvimento de coleções. Evans e Saponaro (2005) propuseram nove ações: identificação, seleção, aquisição, organização, preparação, armazenamento, interpretação, utilização e disseminação. De modo consonante, Johnson (2004) estabeleceu os seguintes procedimentos: avaliação das necessidades dos usuários; seleção; estudos de uso de coleções; análise de coleção; administração de orçamento; identificação das necessidades de cobrança; alcance e ligação com a comunidade e os usuários; planejamento para o compartilhamento de recursos; decisões sobre limpeza, armazenamento e preservação; e a organização e a atribuição de responsabilidades. Vignau e Meneses (2005) seguiram abordagem similar, indicando sete ações: identificação da comunidade, análise da comunidade, política de coleção, política de seleção, aquisição, política de descarte e avaliação da coleção.

Nota-se que, em todas as propostas, o processo é iniciado com o estudo das necessidades do usuário, cujas demandas informacionais são levantadas. Em seguida,

[1] Disponível em: https://github.com.

[2] Disponível em: http://www.periodicos.capes.gov.br.

[3] Disponível em: https://www.rnp.br/servicos/servicos-avancados/cafe.

[4] Disponível em: https://gitlab.com/explore.

[5] Representa um recurso para copiar um projeto em uma determinada plataforma.

[6] Constitui o envio de código de um usuário para determinado projeto.

é realizada a identificação das obras que atenderão tal demanda, complementada pela seleção e pela aquisição. Com a coleção formada, procede-se à organização e ao armazenamento, por conseguinte, manifesta-se a divulgação para o usuário e a coleção passa a ser utilizada. Percebe-se que, para cada etapa do desenvolvimento de coleções, há peculiaridades. Por esse motivo é importante a PDC, pois ela guia o processo, tornando-o, assim, transparente, seguro e colaborativo.

4.2 *Blockchain* e os contratos inteligentes

A *blockchain* possui blocos contendo diversas transações (NAKAMOTO, 2008). Um bloco válido ocorre quando todas as transações estão válidas. O bloco recém validado aponta para o último bloco válido, formando a cadeia de blocos (*blockchain*). Essa cadeia é garantida por algoritmos de criptografia.

Os contratos inteligentes são códigos executáveis que rodam sobre a *blockchain* (ALHARBY; VAN MOORSEL, 2017). De acordo com Zheng *et al*. (2020), os contratos inteligentes, depois de implantados na *blockchain*, mediante uma taxa, são autônomos. Isso quer dizer que não é necessária uma autoridade central, uma vez que os dados necessários estão na própria *blockchain*. Ao assinar um contrato inteligente, as partes envolvidas estão cientes das consequências, não sendo possível reverter a ação. Isso é exequível, pois, assim como todos os dados, os contratos inteligentes estão disponíveis publicamente para auditoria.

Para Xavier e Duque (2021), os contratos inteligentes conseguem armazenar apenas o livro razão (*ledger*) na *blockchain*. Essa ocorrência implica que os servidores externos armazenem quantidades maiores de dados e seus respectivos metadados.

Na triagem inicial dos repositórios públicos de código, foram identificados diversos projetos para a criação de uma *blockchain*, a saber: *Ethereum* Na triagem inicial dos repositórios públicos de código, foram identificados diversos projetos para a criação de uma *blockchain*, a saber: *Ethereum*[7], *Hyperledger Fabric*[8], *Chia*[9], *Solana*[10] e *Flare*[11].

Ainda nos repositórios públicos, foram encontradas linguagens para a criação de contratos inteligentes, como: *Solidity*[12], *Vyper*[13] e *Pact*[14]. Por fim, foram recuperadas bibliotecas para a criação de contratos inteligentes: *Open Zeppelin*[15], *Waffle*[16] e *SolidState*[17]. Ressalta-se que as linguagens são abstrações que estimulam a criação da lógica nos contratos inteligentes, ao passo que as bibliotecas fornecem os códigos comuns que podem ser reaproveitados por diversos contratos.

Cada *blockchain* será submetida a uma avaliação objetiva dos recursos existentes. Com efeito, será escolhida a *blockchain* que ofereça maior quantidade de recursos e que possua suporte da comunidade e dos mantenedores.

4.3 Aplicação dos contratos inteligentes

Uma das descobertas relatadas na literatura é que os contratos inteligentes não são aplicáveis a todo o desenvolvimento de coleções, visto que o fator humano e a gestão do material físico, por exemplo, não possuem uma solução totalmente digital. Nesse contexto, os estudos analisados focam na utilização tecnológica da *blockchain* e nos contratos inteligentes em etapas específicas do desenvolvimento de coleções.

Os contratos inteligentes permitem a criação de papéis para cada parte interessada nas transações (CAO; YANG, 2019). Essa divisão de papéis permite melhorar a análise das transações,

[7] Disponível em: https://github.com/ethereum/go-ethereum.
[8] Disponível em: https://github.com/hyperledger/fabric.
[9] Disponível em: Disponível em: https://github.com/Chia-Network/chia-blockchain.
[10] Disponível em: https://github.com/solana-labs/solana.
[11] Disponível em: https://gitlab.com/flarenetwork/flare.
[12] Disponível em: https://github.com/ethereum/solidity.
[13] Disponível em: https://github.com/vyperlang/vyper.
[14] Disponível em: https://github.com/kadena-io/pact.
[15] Disponível em: https://github.com/OpenZeppelin/openzeppelin-contracts.
[16] Disponível em: https://github.com/TrueFiEng/Waffle.
[17] Disponível em: https://github.com/solidstate-network/solidstate-solidity.

autorizando a interação com os contratos. Assim, são considerados quatro agentes, tais como: autor (pessoas ou organizações que criam a entrada de um item da coleção), editor (auxilia o autor na promoção e na cobrança pela obra), leitor (pessoas ou organizações que alugariam ou pegariam emprestados os itens da coleção) e publicador (agente que publica a cópia ou parte de uma obra). Na visão de Coghill (2018), a identificação dos papéis permite o patrocínio entre essas partes sem intermediação de um administrador central, por exemplo, patrocínio de leitores a um determinado autor.

O direito autoral controlado pela *blockchain* permite que os editores e autores possam receber melhor *feedback* da circulação da obra (HOY, 2017; VERMA, 2021). Ademais, é possível garantir a comissão por transações realizadas, de modo a estimular, com eficiência, a cadeia de produção.

A criação de registros na *blockchain* de cada item da coleção garante o benefício da rastreabilidade, com o propósito de proteger os materiais contra a adulteração e contra a perda de dados (XAVIER; DUQUE, 2021). Nesse sentido, podem ser analisados os itens ociosos, as compras, as vendas e os empréstimos (ZENG *et al*., 2019). Liu (2019) propôs um sistema completo de gestão de itens de uma coleção, cuja aquisição e o empréstimo estão integrados, utilizando tecnologias como: plataformas web, aplicativos móveis e portais. Ademais, os contratos inteligentes podem auxiliar na gestão da submissão e na revisão (CASINO; DASAKLIS; PATSAKIS, 2019). Essa abordagem garante maior transparência nas participações realizadas e as contribuições apresentadas podem ser utilizadas nos processos de identificação, seleção, aquisição e desseleção, durante o desenvolvimento de coleções. Essas contribuições se justificam, uma vez que o processo de registro na *blockchain* gera métricas para nortear a identificação das obras com maiores transações.

Por meio da análise da literatura científica, este estudo não localizou propostas com implementação prática. Também não foi encontrado repositório público de código alinhado à proposta desta pesquisa. De qualquer modo, os bibliotecários precisam acompanhar o desenvolvimento da *blockchain*, sobretudo as operações complexas dos bastidores, analisando, com detalhes, "[...] quando, onde, por que e como a *blockchain* pode ser usada para melhorar a eficiência, eficácia e confiabilidade de dados intensivos" (FREDERICK, 2019, p. 4).

Pelos resultados, nota-se que a temática não se esgotou na revisão da literatura científica, tampouco, na análise de repositórios públicos de código. No entanto, as descobertas iniciais recuperaram relevantes estudos e tecnologias a serem pesquisados com mais profundidade.

5 Considerações finais

Quando se iniciou esta pesquisa, identificou-se uma nova abordagem para o desenvolvimento de coleções, sendo esse um processo fundamental para o crescimento de uma biblioteca. Entretanto, a *blockchain* e os contratos inteligentes não foram aproveitados em sua potencialidade para auxiliar no gerenciamento das coleções.

A metodologia utilizada foi a revisão da literatura, tendo em vista mapear os estudos em que a *blockchain* e os contratos inteligentes foram adotados no âmbito do desenvolvimento de coleções. Em seguida, foi realizada a análise dos repositórios públicos para identificação da *blockchain* e dos contratos inteligentes.

A pesquisa está em andamento e não esgotou as revisões. O objetivo geral, que é identificar as contribuições dos contratos inteligentes no desenvolvimento de coleções, foi alcançado, de modo parcial. Já os objetivos específicos têm potencial de pesquisa, cujas descobertas garantirão robustas contribuições para a área da Biblioteconomia e Ciência da Informação.

Pelo estado atual do estudo, considera-se que o problema está respondido, em parcial, pois foram identificadas muitas pesquisas teóricas, mas nenhuma implementação. Assim, haverá a necessidade da criação específica da *blockchain* e dos contratos inteligentes para o desenvolvimento de coleções.

Uma limitação desta pesquisa é não priorizar questões de segurança e

privacidade. Isso indica a necessidade da realização de estudos posteriores.

Referências

ALHARBY, M.; VAN MOORSEL, A. Blockchain-based smart contracts: A systematic mapping study. **arXiv preprint arXiv:1710.06372**, 2017.

CAO, L.; YANG, H.. Building virtual digital library based on P2P and Blockchain. *In*: **2019 11th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)**. IEEE, 2019. p. 341-345.

CASINO, F.; DASAKLIS, T. K.; PATSAKIS, C. A systematic literature review of blockchain-based applications: Current status, classification and open issues. **Telematics and informatics**, v. 36, p. 55-81, 2019

GIL, A. C. Como elaborar projetos de pesquisa. 4. ed. São Paulo: Atlas, 2002.

COGHILL, Jeffrey G. Blockchain and its implications for libraries. **Journal of Electronic Resources in Medical Libraries**, v. 15, n. 2, p. 66-70, 2018.

EVANS, G. E.; SAPONARO, M. Z. **Developing library and information center collections**. 5. ed. Westport: Libraries Unlimited, 2005.

FREDERICK, D. E. Blockchain, libraries and the data deluge. **Library Hi Tech News**, v. 36, n. 10, p. 1-6, 2019.

HOY, M. B. An introduction to the blockchain and its implications for libraries and medicine. **Medical reference services quarterly**, v. 36, n. 3, p. 273-279, 2017.

JOHNSON, P. **Fundamentals of collection development & management**. Chicago: ALA, 2004.

LIU, X. A smart book management system based on Blockchain platform. *In*: **2019 International Conference on Communications, Information System and Computer Engineering (CISCE)**. IEEE, 2019. p. 120-123.

NAKAMOTO, S. Bitcoin: A peer-to-peer electronic cash system. **Decentralized Business Review**, p. 1-9, 2008.

REITZ, J. M. **Dictionary for library and information science**. Libraries Unlimited, 2004.

SAFDAR, M. *et al*. A mapping review of literature on Blockchain usage by libraries: Challenges and opportunities. **Journal of Librarianship and Information Science**, p. 1-11, 2022.

SANJAY, P. Collection development in academic libraries. **International Journal of Library and Information Science**, v. 8, n. 7, p. 62-67, 2016.

VERGUEIRO, W. C. S. Desenvolvimento de coleções: uma nova visão para o planejamento de recursos informacionais. **Ciência da Informação**, Brasília, v. 22, n. 1, 1993.

VERMA, M. Amalgamation of Blockchain Technology and Knowledge Management System to fetch an enhanced system in Library. **International Journal of Innovative Research in Technology**, v. 7, n. 11, p. 474-477, 2021.

VIGNAU, B. S. S.; MENESES, G.. Collection development policies in university libraries: a space for reflection. **Collection building**, 2005.

WEITZEL, S. R. Desenvolvimento de coleções: origem dos fundamentos contemporâneos. **Transinformação**, v. 24, p. 179-190, 2012.

XAVIER, A. C. C.; DUQUE, C. G. Prontuário eletrônico do paciente: qual a contribuição da arquivística e do Smart Contracts para a sua gestão na Era da Saúde 4.0?. **AtoZ: novas práticas em informação e conhecimento**, v. 10, n. 3, p. 1-10, 2021.

ZENG, J. *et al*. BookChain: Library-free book sharing based on Blockchain technology. *In*: **2019 15th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN)**. IEEE, 2019. p. 224-229.

ZHENG, Z. *et al*. An overview on smart contracts: Challenges, advances and platforms. **Future Generation Computer Systems**, v. 105, p. 475-491, 2020.

CRIAÇÃO E CAPTURA DE VALOR BASEADAS EM BIG DATA PARA A INOVAÇÃO EM PRODUTOS E SERVIÇOS: ANÁLISE DA PRODUÇÃO CIENTÍFICA

BIG DATA BASED ON VALUE CREATION AND VALUE CAPTURE FOR INNOVATION IN PRODUCTS AND SERVICES: AN ANALYSIS OF SCIENTIFIC PRODUCTION

Priscila Machado Borges Sena[1], Nathalia Berger Werlang[2], Ana Clara Cândido[3],
(1) Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), SAUS Quadra 5 - Lote 6, Bloco H, Brasília – DF, 70070-912, priscilasena@ibict.br
(2) Departamento de Ciência da Informação (CIN), Universidade Federal de Santa Catarina (UFSC), Campus Universitário Reitor João David Ferreira Lima, Florianópolis – SC, 88040-900, nathalia.werlang@ufsc.br
(3) Programa de Pós-graduação em Ciência da Informação (PGCIN), Departamento de Ciência da Informação (CIN), Universidade Federal de Santa Catarina (UFSC), Campus Universitário Reitor João David Ferreira Lima, Florianópolis – SC, 88040-900, ana.candido@ufsc.br

Resumo
A informação é um recurso essencial para o desenvolvimento de diversas atividades organizacionais. Neste sentido, a criação e captura de valor por meio do *Big Data* é atividade estratégica na atual gestão das organizações. Sendo assim, esta pesquisa tem como objetivo reconhecer na literatura científica abordagens que relacionam a criação e captura de valor baseadas em *Big Data* para a inovação em produtos e serviços no âmbito dos negócios. Os procedimentos metodológicos adotados para a coleta de dados seguiram os princípios da busca sistemática da literatura nas bases de dados Scopus e Web of Science (WoS). Ao final, foram selecionados e analisados 18 artigos de maneira qualitativa. Os principais resultados apontam que a captura de informação de diferentes parceiros, clientes e usuários por meio da tecnologia da informação permite às organizações a criação de valor, resolução de problemas e inovação organizacional. É possível concluir que a temática ainda é emergente nos estudos organizacionais e novas investigações são necessárias especialmente para compreender o fenômeno em países emergentes.

Palavras-chave: Informação; Criação de valor; Captura de valor; Big Data; Inovação.

Abstract
Information is an essential resource for the development of various organizational activities. In this sense, the creation and capture of value through Big Data is a strategic activity in the current management of organizations. Thus, this research aims to recognize in the scientific literature approaches that relate the creation and capture of value based on Big Data for innovation in products and services in the business environment. The methodological procedures adopted for data collection followed the principles of a systematic literature search in the Scopus and Web of Science (WoS) databases. In summary, 18 articles were selected and analyzed qualitatively. The main results indicate that capturing information from different stakeholders, customers, and users through information technology enables organizations to create value, solve problems, and innovate organizationally. It is possible to conclude that the theme is still emerging in organizational studies and further research is needed primarily to understand the phenomenon in emerging countries.

Keywords: Information; Value Creation; Value Capture; Big Data; Innovation.

1 Introdução

A informação como recurso está intrínseca em diversas atividades organizacionais, embora nem sempre as organizações se atentem para a importância do processo de gestão da informação. Na rotina organizacional a informação é suporte, por exemplo, nos produtos e serviços informacionais (sistemas, relatórios, dashboards) que fornecem dados e informação para a tomada de decisão em esferas diversas (introdução de um novo produto/ serviço, aquisição/instalação de uma unidade organizacional etc.).

De acordo com Urbinati et al. (2018), a necessidade de sistematização da informação é uma lacuna nas pesquisas de *Big Data* quando se remete a Transformação Digital, especificamente na identificação das necessidades dos clientes; gestão de risco e tomada de decisão; conhecimento orientado a dados; design de produtos e serviços; gestão da qualidade; criação e reconhecimento de oportunidade.

Conforme Davenport (2017), o mais imprescindível na tecnologia de *Big Data* é como ela pode criar valor para a sua organização: diminuir os custos e impulsionar a velocidade do processamento de dados; contribuir no desenvolvimento de novos produtos ou serviços; ou subsidiar melhor suporte, como novos dados e modelos, ao processo decisório.

Pressupõe-se que a partir da criação e captura de valor baseadas em *Big Data*, as organizações podem se tornar mais assertivas em suas tomadas de decisões, e ainda, mais inovadoras. Uma vez que, compreende-se que hoje a inovação deve ser uma constante para que as organizações consigam se diferenciar e tornarem-se cada vez mais competitivas.

Posto isso, este trabalho resulta da 1ª fase do "Projeto Criação e captura de valor no desenvolvimento de novos produtos e serviços baseadas em *Big Data*: análise de empresas catarinenses", apoiado pela Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC), Edital nº 09/2022 (SANTA CATARINA, 2022).

Na primeira fase se busca realizar um levantamento sistemático da literatura sobre captura de valor a partir de *Big Data* para identificar o estado da arte, evolução e tendências de pesquisa sobre a temática, em consonância com o subtópico "Impacto e Evolução" no tópico "Big Data", presente no Workshop de Informação, Dados e Tecnologia (WIDaT 2022), ao qual se submete este trabalho.

2 Objetivos

Diante do contexto apresentado na introdução deste trabalho, objetiva-se reconhecer na literatura científica abordagens que relacionam a criação e captura de valor baseadas em *Big Data* para a inovação em produtos e serviços no âmbito dos negócios. Ademais, visa-se identificar métodos e instrumentos que auxiliem no processo de coleta e análise dos dados.

3 Procedimentos Metodológicos

Compreende-se a pesquisa desenvolvida por exploratória e descritiva, com abordagem qualitativa. Para a concretização do objetivo proposto neste trabalho por intermédio do Portal de Periódicos da CAPES, realizou-se buscas sistemáticas na Scopus, na Web of Science (WoS) e em todas as bases de dados de sua coleção, com os termos (("value capture" AND "Big Data")) no título, resumo e palavras-chave dos documentos, sem qualquer tipo de filtro de data, área ou tipo de documento. As bases utilizadas foram escolhidas por oferecerem uma cobertura mundial da produção científica.

Encontrou-se 15 documentos na Scopus e 23 na WoS. Eliminadas as duplicações, a busca resultou em 24 documentos. Excluiu-se 4 documentos referentes a anais completos de eventos e editorial.

Após a leitura dos resumos para identificação do alinhamento ao escopo deste trabalho, selecionou-se 18 dos 20 documentos.

4 Resultados

O portfólio pertinente ao objetivo da pesquisa descrita se constituí de artigos situados no período temporal de 2017 a 2022. Destaca-se os três últimos anos com maior número de publicações (2020 = 5, 2021 = 4, 2022 = 4), apresentando abordagens que relacionam a criação e captura de valor baseadas em Big Data para a inovação em produtos e serviços no âmbito dos negócios. Em sua maioria pesquisas empíricas (12), seguidas de pesquisas teóricas (4), e teóricas-empíricas (2).

Em ordem cronológica crescente, inicia-se por Ramkumar et al. (2017), que abordaram a parceria público-privada para o desenvolvimento das práticas de medicina e saúde pública suportada por dispositivos móveis (mHealth). Discorreram sobre a arquitetura aberta como oportunidade a ser explorada na captura de valor baseada em *Big Data*.

Mamonov e Triantoro (2018) desenvolveram pesquisa voltada às indústrias emergentes. Utilizaram o método de estudo de caso, com a análise de dois casos representativos de empresas de pequeno e médio porte, dos nichos de financiamento e publicidade online. Dos resultados, verificaram que as empresas utilizaram a vantagem inicial obtida através de novas aplicações de fontes de dados e tecnologia de informação para desenvolver serviços padronizados de tecnologias de informação por uma grande rede de clientes

e parceiros; observaram que enquanto os investimentos em recursos de dados podem permitir a criação de valor por meio de capacidades analíticas únicas, o potencial de dissociação entre a criação de valor e a captura de valor pode impedir que as empresas possam colher todo o valor criado através desses recursos de dados.

Brinch (2018) buscou compreender o valor de *Big Data* na gestão da cadeia de suprimentos, por meio de revisão sistematizada da literatura científica. Evidenciou, portanto, que a adoção de *Big Data* na gestão da cadeia de suprimentos representa uma maior ênfase nos dados como um *input* para um processo do que foi considerado anteriormente. Além disso, destacou que a literatura científica em torno da gestão da cadeia de suprimentos carece de entendimento aprofundado acerca de *Big Data* e do seu valor.

Urbinati et al. (2018) exploraram como as empresas provedoras criam e capturam valor a partir de *Big Data*, baseando-se em um estudo de caso múltiplo de empresas provedoras que oferecem soluções e serviços baseados em *Big Data*. Os resultados ilustraram uma estrutura teórica sobre criação e captura de valor, confiando no *Big Data* e identificando duas principais estratégias de serviços de inovação baseadas em *Big Data* utilizadas pelas empresas provedoras.

Berawi et al. (2019) apresentaram pesquisa da aplicação da mineração de dados para determinar as variáveis que afetam a instabilidade do preço do imóvel e a correlação com a proximidade da estação de trânsito. Destacaram que quanto mais próximo de uma propriedade da estação de trânsito, o preço seria duas vezes mais barato em comparação com aqueles localizados mais distantes.

Wiener et al. (2020) abordaram a renovação dos modelos de negócios ou desenvolvimento novos modelos, dando origem ao fenômeno dos modelos de negócios de *Big Data*.

Pettit et al. (2020) discorreram sobre a era digital *Big Data*, análise de dados e cidades inteligentes, para elucidar uma nova geração de sistemas de apoio ao planejamento. O Rapid Analytics Interactive Scenario Explorer foi o sistema detalhado, uma vez que visa apoiar o planejamento desenvolvido para ajudar os planejadores e formuladores de políticas a determinar o provável aumento do valor do terreno associado ao fornecimento de nova infraestrutura urbana.

Elia et al. (2020) apresentaram a importância da inovação aberta, a partir da exemplificação da relevância de uma comunidade virtual de marcas como nova ferramenta de gestão para inovação aberta. As pesquisas existentes sobre a interseção entre as comunidades de marcas e a gestão da inovação mostraram como o conhecimento dos membros das comunidades de marcas e seu envolvimento em discussões relacionadas aos produtos representam uma fonte relevante de inovação para as empresas. Entretanto, como as empresas empregam a inovação aberta por meio das comunidades virtuais de marcas e como elas implementam práticas intencionais tornam-se uma lacuna de pesquisa inexplorada.

Canhoto e Clear (2020) refletiram sobre o fato da inteligência artificial (IA) e a aprendizagem de máquinas (do inglês machine learning - ML) poderem economizar dinheiro e melhorar a eficiência dos processos comerciais, mas também poderem destruir o valor comercial, às vezes com graves consequências. Propuseram então uma nova estrutura para mapear os componentes de uma solução de IA e para identificar e gerenciar o potencial de destruição de valor da IA e ML para as empresas. Mostraram como as características que definem a IA e o ML podem ameaçar a integridade das entradas, processos e resultados do sistema de IA. Em seguida, extraíram dos conceitos de conteúdo de criação de valor e processo de criação de valor para mostrar como esses riscos podem dificultar a criação de valor ou mesmo resultar na destruição de valor. Por fim, ilustraram a aplicação de da estrutura com um exemplo de implementação de um chatbot alimentado por IA no atendimento ao cliente, e discutiram como solucionar os problemas que surgem.

Berawi et al. (2020) descreveram a utilização de *Big Data* para acelerar o processamento de dados de pesquisa empregando o método de mineração de

dados, bem como o sistema de informação geográfica (GIS) e a modelagem de preços hedônicos (HPM) para investigar os preços de propriedades a fim de formular evidências empíricas para a pesquisa. Os resultados mostraram evidências opostas em comparação com estudos anteriores que argumentavam que a acessibilidade poderia contribuir para o preço da propriedade quando mais perto da estação de trânsito para o desenvolvimento de mais pesquisas.

Klos et al. (2021) buscaram compreender como a transformação do modelo de negócios digital pode ser alcançada, por meio do estudo de 15 casos. As conclusões se concentraram em dados de entrevistas e registros de arquivos complementares de 15 casos. Apresentaram uma estrutura para a transformação do modelo de negócios digital ao longo das dimensões da proposta de valor, criação de valor e captura de valor. Enfatizaram a importância de uma fase preparatória na qual o curso estratégico é definido. Demonstraram que a transformação do modelo de negócios de uma empresa é mais eficaz quando uma única pessoa, ou seja, o Diretor-Geral Digital, é responsável. Por fim, contribuíram para a literatura do modelo de negócios, fornecendo uma visão mais holística sobre como a inovação do modelo de negócios pode ser utilizada durante a transformação digital.

Ammirato et al. (2021) propuseram detectar modelos de negócios e características-chave dos aplicativos móveis para o turismo cultural e analisar a oferta de serviços baseados em aplicativos neste setor. Os autores definiram uma metodologia para identificar, caracterizar e analisar uma categoria particular de produtos digitais para o turismo cultural: serviços baseados em aplicativos. Eles são estudados em termos de criação de valor, proposição e captura com o objetivo de identificar as características distintivas dos modelos de negócios. Como resultado, os autores identificaram uma estrutura de classificação em três dimensões principais, a saber "como explorar recursos de aplicativos móveis para criar valor para turistas culturais" (criação de valor), "quais serviços valiosos são prestados a turistas culturais" (proposta de valor) e "como as empresas são recompensadas pelo valor que oferecem" (captura de valor). Tal estrutura representa uma ferramenta prática que fornece insights frutíferos para o projeto de uma nova geração de serviços baseados em aplicativos dentro do chamado domínio "Internet das coisas".

Cappa et al. (2021) discutiram sobre a escassez de pesquisas sobre o impacto de *Big Data* no desempenho das empresas a longo prazo. Focalizaram em como as empresas criam e capturam valor a partir de *Big Data* sobre clientes, com a visão baseada em recursos e três dimensões de *Big Data* (volume, variedade e veracidade) para entender quando os benefícios superam os custos. Ao checarem o número de downloads de aplicações de dispositivos móveis, descobriram que o volume de Big Data tem um efeito negativo sobre o desempenho da empresa. Sugeriram que o "*bigness*" de Big Data por si só não garante a criação de valor para uma empresa, e poderia até mesmo constituir um "lado negro" Big Data.

Favoretto et al. (2021), por meio da revisão sistemática de literatura, visaram identificar os desafios da transformação digital nas empresas de manufatura e propor novas direções de pesquisa. Identificaram uma visão sistematizada dos desafios relativos ao compromisso organizacional, criação de valor, proposta de valor, entrega de valor, captura de valor, infraestrutura de informação e tecnologia e segurança de dados. Além disso, desenvolveram uma estrutura conceitual para resumir os desafios e como eles estão associados com a arquitetura de valor do modelo de negócios e com as fases da transformação digital.

Temiz et al. (2022) investigaram as razões subjacentes ao investimento em dados abertos. Com base nos resultados da pesquisa, entrevistas e provas complementares de fontes secundárias, exploraram os motivos e crenças sobre o investimento em dados abertos expressos por especialistas em dados abertos, tanto em organizações públicas quanto privadas. Descobriram que tanto em organizações públicas quanto privadas, os investimentos em dados abertos são impulsionados mais pela busca de legitimidade do que por uma busca para realizar o potencial de criação de valor dos dados abertos.

Loonam e O'Regan (2022), por meio da análise da literatura científica, elucidaram

que de particular relevância para as cadeias de valor globais é o surgimento de tecnologias digitais, como o *Big Data*, a Internet das Coisas e a inteligência artificial. Oito temas-chave emergiram da literatura para revelar questões críticas para as plataformas digitais. As implicações para a estratégia são então discutidas, revelando quatro aprendizados-chave para as cadeias de valor globais, a saber, (i) gestão de fronteiras, (ii) definição da unidade de análise, (iii) alinhamento de capacidades, e (iv) governança e liderança.

Cuomo et al. (2022) propuseram uma abordagem baseada em dados para impulsionar a experiência turística em serviços integrados de mobilidade e discutir como a experiência pode ser melhorada. Em particular, a abordagem orientada por dados, devido ao projeto de um sistema de recomendação baseado em um mecanismo de análise de dados de grande porte, torna possível: i) classificar as preferências turísticas para os destinos italianos mais atraentes no Google; ii) classificar as principais atrações - lazer, entretenimento, cultura etc. - associados a destinos turísticos únicos, obtidos a partir da análise de sites temáticos relevantes, como Tripadvisor, Minube e Travel365. O estudo dependeu do apoio de Big Data social para o conceito de co-design de experiência turística, com foco em serviços integrados de mobilidade. Do ponto de vista tecnológico, a análise de Big Data é possibilitada por contar com uma plataforma de dados baseada em nuvem, como o Amazon web services (AWS), Microsoft Azure ou a plataforma de nuvem do Google (GCP). Isto provou ser a chave para coletar, atualizar e processar regularmente dados de várias fontes heterogêneas, tais como consultas de busca do Google acessíveis via Google Trends, ou qualquer dado social raspado de websites, bem como extrair insights relevantes que possam atender às necessidades comerciais expressas pelas empresas de mobilidade.

Browder et al. (2022), associaram diretamente a transformação digital com "análises de *Big Data*" (BDA). Ao analisar uma relação interorganizacional na qual tanto o fornecedor quanto o cliente tentam desenvolver novos produtos e serviços com BDA, esta pesquisa examina as condições sob as quais BDA leva as organizações a explorar novos processos de inovação. A pesquisa fornece uma compreensão contextual da criação de valor com BDA nas relações intraorganizacionais e contribui para a literatura sobre inovação e transformação digital com tecnologias emergentes.

Ante a análise dos artigos, identificou-se métodos e instrumentos para auxílio no processo de coleta e análise dos dados em Urbinati et al. (2018), Browder et al. (2022), Klos et al. (2021), Loonam e O'Regan (2022) e Temiz et al. (2022).

5 Considerações Finais

Ao avaliar os achados desta pesquisa, conclui-se que o objetivo proposto foi alcançado, já que se tornou possível reconhecer na literatura científica abordagens que relacionam a criação e captura de valor baseadas em *Big Data*. A busca sistemática de literatura permitiu a leitura e análise de 18 estudos que trouxeram importantes insights para o desenvolvimento da temática.

Inicialmente, destaca-se que a captura de informação de diferentes parceiros, clientes e usuários por meio da tecnologia da informação permite às organizações a criação de valor, resolução de problemas e inovação organizacional.

Os estudos ainda elucidaram que a cocriação de valor ocorre por meio da integração e captura desses dados de diferentes atores, que juntos, promovem a melhoria dos serviços e produtos ofertados pelas organizações.

As pesquisas também apontam que o uso de *Big Data* permite a compreensão, identificação e criação de produtos ou serviços mais assertivos e significativos para o público-alvo, uma vez que a obtenção e análise dos dados dos usuários aponta suas demandas e necessidades em tempo real.

Assim, as organizações que tomam decisões baseadas em *Big Data*, obtêm maior vantagem competitiva em relação aos seus concorrentes, já que cocriam, inovam e geram valor percebido aos seus usuários.

Por fim, esta pesquisa ainda aponta a possibilidade de utilizar dados abertos para criação de valor por meio da utilização de ferramentas, a exemplo de consultas de busca do Google.

Como sugestão de futuras pesquisas, reforça-se a continuidade das investigações sobre *Big Data,* especialmente relacionada com as temáticas: inovação, cocriação e ciência aberta. Todos os setores da economia podem se beneficiar com a utilização de *Big Data* e são necessários novos estudos especialmente em países emergentes.

Agradecimentos

À Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC) pelo subsídio financeiro concedido.

Referências

AMMIRATO, Salvatore *et al*. Digital business models in cultural tourism. **International Journal of Entrepreneurial Behavior & Research**, 2021.

BERAWI, Mohammed Ali *et al*. Impact of rail transit station proximity to commercial property prices: utilizing Big Data in urban real estate. **Journal of Big Data**, v. 7, n. 1, p. 1-17, 2020.

BERAWI, Mohammed Ali *et al*. Land value capture modeling in commercial and office areas using a Big Data approach. **International Journal of Technology**, v. 10, n. 6, p. 1150-1150, 2019.

BRINCH, Morten. Understanding the value of Big Data in supply chain management and its business processes: Towards a conceptual framework. **International Journal of Operations & Production Management**, 2018.

BROWDER, Russell E. *et al*. Learning to Innovate with Big Data Analytics in Interorganizational Relationships. **Academy of Management Discoveries**, v. 8, n. 1, p. 139-166, 2022.

CANHOTO, Ana Isabel; CLEAR, Fintan. Artificial intelligence and machine learning as business tools: A framework for diagnosing value destruction potential. **Business Horizons**, v. 63, n. 2, p. 183-193, 2020.

CAPPA, Francesco *et al*. Big Data for creating and capturing value in the digitalized environment: Unpacking the effects of volume, variety, and veracity on firm performance. **Journal of Product Innovation Management**, v. 38, n. 1, p. 49-67, 2021.

CUOMO, Maria Teresa *et al*. Enhancing Traveller Experience In Integrated Mobility Services Via Big Social Data Analytics. **Technological Forecasting and Social Change**, v. 176, p. 121460, 2022.

DAVENPORT, Thomas H. **Big Data no trabalho:** derrubando mitos e descobrindo oportunidade. Rio de Janeiro: Alta Books, 2017.

ELIA, Gianluca; PETRUZZELLI, Antonio Messeni; URBINATI, Andrea. Implementing open innovation through virtual brand communities: A case study analysis in the semiconductor industry. **Technological forecasting and social change**, v. 155, p. 119994, 2020.

FAVORETTO, Camila *et al*. Digital transformation of business model in manufacturing companies: challenges and research agenda. **Journal of Business & Industrial Marketing**, 2021.

JABŁOŃSKI, Marek. Value migration to the sustainable business models of digital economy companies on the capital market. **Sustainability**, v. 10, n. 9, p. 3113, 2018.

KLOS, Christoph *et al*. Digital transformation of incumbent firms: a business model innovation perspective. **IEEE Transactions on Engineering Management**, 2021.

LOONAM, John; O'REGAN, Nicholas. Global value chains and digital platforms: Implications for strategy. **Strategic Change**, v. 31, n. 1, p. 161-177, 2022.

MAMONOV, Stanislav; TRIANTORO, Tamilla Mavlanova. The strategic value of data resources in emergent industries. **International Journal of Information Management**, v. 39, p. 146-155, 2018.

PETTIT, Chris *et al*. A new toolkit for land value analysis and scenario planning. **Environment and Planning B: Urban Analytics and City Science**, v. 47, n. 8, p. 1490-1507, 2020.

RAMKUMAR, Prem N. *et al*. Open mHealth architecture: a primer for tomorrow's orthopedic surgeon and introduction to its use in lower extremity arthroplasty. **The Journal of arthroplasty**, v. 32, n. 4, p. 1058-1062, 2017.

SANTA CATARINA. Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC). **Edital de Chamada Pública Fapesc n° 09/2022 – Apoio à Pesquisa Aplicada sobre Complexidade Econômica, Inovação e Prioridades para o Desenvolvimento Estadual Catarinense**. Florianópolis, 15 mar. 2022.

TEMIZ, Serdar *et al*. Open data: Lost opportunity or unrealized potential?. **Technovation**, p. 102535, 2022.

URBINATI, Andrea *et al*. Creating and capturing value from Big Data: A multiple-case study analysis of provider companies. **Technovation**, v. 84, p. 21-36, 2018.

WIENER, Martin; SAUNDERS, Carol; MARABELLI, Marco. Big-data business models: A critical literature review and multiperspective research framework. **Journal of Information Technology**, v. 35, n. 1, p. 66-91, 2020.

DESCOBERTA DE CONHECIMENTO APLICADA À BASE DE DADOS ABERTOS DA ANVISA SOBRE PREÇOS DE MEDICAMENTOS POR MEIO DE ANÁLISE DE REDES DE INFORMAÇÃO

DISCOVERY OF KNOWLEDGE APPLIED TO THE OPEN DATABASE OF ANVISA ABOUT DRUG PRICES THROUGH INFORMATION NETWORK ANALYSIS

Lucas Vale[1], Henrique Monteiro Cristovão[2]
(1) Universidade Federal do Espírito Santo, Campus de Goiabeiras, lucas.s.vale@edu.ufes.br
(2) Universidade Federal do Espírito Santo, Campus de Goiabeiras, henrique.cristovao@ufes.br

Resumo
Propõe-se nesta pesquisa investigar e revelar possíveis relações entre variáveis da base de dados abertos da Anvisa sobre preço de medicamentos. A pesquisa tem abordagem qualitativa e natureza aplicada. A base de dados analisada é mantida e alimentada pela ANVISA, disponível no portal de dados abertos do governo, sendo composta por 26310 registros correspondentes ao período de 2017 a 2021. Fundamentado na metodologia de descoberta de conhecimento de Fayyad onde, nas fases pré-processamento e transformação, usou-se o software OpenRefine, e nas fases mineração de dados, interpretação e avaliação aplicou-se a análise de redes complexas com suporte dos softwares OpenRefine e Microsoft Power BI Desktop. Observa-se, apesar da ausência de alguns registros, que os medicamentos mais produzidos pelos laboratórios registrados são os de tarja vermelha, onde estão medicamentos que tratam condições de alta prevalência no Brasil, como hipertensão e diabetes, enquanto que os medicamentos de tarja preta têm produção e demanda muito mais limitadas pela suas restrições de uso. Destaca-se a alta produção de classes terapêuticas em que o custo de produção gira em torno de R$100, indicando que a maioria dos laboratórios têm em seu modelo de negócios públicos-alvo das classes C e D, ao passo que apenas um laboratório é responsável pela produção de todos os medicamentos em que o custo de produção ultrapassa 1 milhão de reais. Entende-se ainda que é preciso dedicar mais esforços na análise da base de dados a fim de potencializar a descoberta de novas relações entre variáveis.

Palavras-chave: Descoberta de conhecimento; Anvisa; Análise de redes complexas; Ciência de dados.

Abstract
The research proposes to investigate and reveal possible relationships between variables in Anvisa's open database on drug prices. The research has a qualitative approach and an applied nature. The analyzed database is maintained and fed by ANVISA, available on the government's open data portal, consisting of 26,310 records corresponding to the period from 2017 to 2021. Based on Fayyad's knowledge discovery methodology where, in the pre-processing and transformation, the OpenRefine software was used, and in the data mining, interpretation and evaluation phases, the analysis of complex networks was applied with support from the OpenRefine and Microsoft Power BI Desktop software. It is observed, despite the absence of some records, that the drugs most produced by registered laboratories are those with the red stripe, which contain drugs that treat highly prevalent conditions in Brazil, such as hypertension and diabetes, while the black stripe drugs have production and demand much more limited by their use restrictions. The high production of therapeutic classes stands out, in which the production cost is around R$100, indicating that most laboratories have in their business model target audiences from classes C and D, while only one laboratory is responsible for the production of all medicines in which the production cost exceeds 1 million reais. It is also understood that it is necessary to dedicate more efforts to the analysis of the database in order to enhance the discovery of new relationships between variables.

Keywords: Discovery of knowledge; Anvisa; Complex networks analysis; Data Science.

1 Introdução

É concomitante ao desenvolvimento da civilização o acúmulo de dados onde, ter acesso à informação desejada de maneira precisa é um diferencial que pode fazer prosperar negócios em áreas muito distintas. Isso faz dos profissionais responsáveis pelo processo de tomada de decisão dependentes de ferramentas assertivas para os procedimentos de coleta, organização, processamento e utilização da informação. (COSTA et al., 2009)

Temas relacionados à saúde e ao bem-estar são sempre sensíveis e relevantes para

a sociedade. No contexto onde a população está cada vez mais envelhecida, entende-se que é primária a preocupação com um Sistema Nacional de Saúde eficiente e robusto (BÁRRIOS et al. 2020). No Brasil, o Sistema Único de Saúde (SUS) tem integrado a si a autarquia ANVISA (Agência Nacional de Vigilância Sanitária), responsável por, entre outras coisas, a gestão do controle de qualidade e da vigência de preços de medicamentos.

O processo de descoberta de conhecimentos (KDD - *Knowledge Discovery in Databas*e) é um ponto de partida metodológico para a identificação de padrões em um determinado conjunto de dados em que exista potencial de compreensão e utilidade para o usuário. No processo sugerido por Fayyad et al. (1996), existem 6 etapas: seleção de dados, pré-processamento, transformação, mineração de dados, interpretação e avaliação e implantação do conhecimento.

A etapa de pré-processamento é onde os dados são manipulados com o propósito de solucionar problemas, corrigir erros, modificar a estrutura, adequar tipos entre outros, preparando-os para que a fase de descoberta de conhecimento seja mais eficiente. Nesta etapa pode-se aprender mais sobre os dados em uso, por meio de visualizações. De modo geral, a fase de pré-processamento é tida como semi-automática, isto é, dependente da aptidão do operador em identificar problemas no conjunto de dados e a sua natureza, como expõe Batista (2003).

Na etapa da transformação podem ocorrer alterações significativas sobre os dados e, principalmente, a criação de novos subconjuntos de dados para atender a requisitos específicos de determinados softwares de mineração.

A etapa de mineração de dados tem grande diversidade de métodos disponíveis para uso. Na presente pesquisa há ênfase na técnica de análise de redes complexas por meio de, principalmente, o uso de métricas, comunidades, projeções bipartidas e inspeção visual. Uma rede complexa, segundo Barabási (2003), é um conjunto de nós interligados por arestas que constitui uma estrutura topográfica sofisticada. O estudo das redes neste formato teve início com a solução do problema das pontes de Königsberg por Euler, em 1735, que derivou a teoria dos grafos (METZ, 2007). A análise de redes complexas, portanto, se destina a compreender a relação entre nós e as suas consequências, pouco se preocupando com as características particulares dos nós, e sim com a sua totalidade estrutural, tomando emprestados fundamentos da análise de redes sociais, como sugerem Wasserman et al. (1994). Nesta etapa, os dados dispostos em nós e arestas fornecem possibilidades de inspeção visual, desde que adequadamente formatados sobre a topologia e as relações existentes entre os nós.

A inspeção visual, ou visualização dos dados, é estrutural na análise de redes de informação para tornar padrões despercebidos evidentes (CHEN, 2013). O processo de visualização pode, além do mais, ser constituinte da fase de pré processamento ou de transformação, uma vez que os limites entre as fases delimitadas pelo KDD não são bem estabelecidos, segundo HAND et al. (2001), sendo de toda forma um artifício da mineração sobretudo da interpretação dos dados.

Imagina-se que o conjunto de dados escolhido possa, ao ser analisado, revelar padrões ou tendências não evidentes ou despercebidas que, em função da sua forte relevância social, podem trazer algum ganho para os indivíduos diretamente envolvidos com o preço de medicamentos no Brasil.

2 Objetivos

Investigar e revelar relações entre variáveis da base de dados abertos da ANVISA sobre preços de medicamentos.

3 Procedimentos Metodológicos

A pesquisa tem abordagem qualitativa e natureza aplicada. A base de dados analisada é mantida e alimentada pela ANVISA, disponível no portal de dados abertos do governo[1] de maneira estruturada e é composta por 26310 registros correspondentes ao período de 2017 a 2021, e com tamanho aproximado de 10Mb.

Após a obtenção dos dados as etapas do KDD foram executadas conforme

[1] Disponível em: *https://dados.gov.br/dataset/preco-de-medicamentos-no-brasil-consumidor*

descritas nas próximas subseções e ilustradas na Figura 1.

3.1 Pré-processamento

Nesta etapa foram removidos ruídos da base de dados e selecionados os atributos mais relevantes para a análise, bem como a formatação adequada dos dados.

Fazendo uso da ferramenta OpenRefine[2], foram excluídas as colunas CNPJ, Registro, EAN, PF 0%, PF 12%, PF 17,5%, PF 17,5% ALC, PF 18%, PF 18% ALC, PF 20%, PMC 12%, PMC 17,5%, PMC 17,5% ALC, PMC 18%, PMC 18% ALC, PMC 20%, Código GGREM, por serem irrelevantes para o estudo, do ponto de vista dos autores e objetivos da pesquisa. O nome das colunas foi alterado para o padrão camelCase[3] para melhor integração entre as ferramentas de análise utilizadas. As colunas contendo valores numéricos foram convertidas para a classe de números (variável quantitativa contínua) e os valores coluna de *tarja,* que continham irregularidades derivadas da inserção dos registros foram padronizados em 4 categorias: *Tarja Preta, Tarja Vermelha, Venda Livre e Sem Tarja*.

3.2 Transformação

Neste ponto, alguns dados são transcritos para formatos mais adequados para a análise, considerando-se peculiaridades do método de análise de redes de informação. Utilizando a linguagem GREL[4], os valores numéricos de preço (quantitativos contínuos) foram transformados em categorias de preço (qualitativos ordinais) seguindo a seguinte organização para as categorias:

- Abaixo de R$100
- Entre R$100 e R$1000
- Entre R$1000 e R$50000
- Entre R$50000 e 1 milhão
- Acima de 1 milhão de reais

Em seguida, foram geradas redes de informação por meio de mapeamento realizado pelo software OpenRefine para o formato de rede GML[5] reconhecido pelo software Gephi[6]. Como exemplo, consta na Figura 2 do Apêndice A o mapeamento realizado para a criação da rede tripartida citada na seção 3.3.

3.3 Mineração de dados, interpretação e avaliação

Na fase de mineração utilizou-se a metodologia de análise de redes complexas que, no caso da presente pesquisa, foram redes de informação. Também foi criado um *dashboard* para a visualização sintetizada de alguns dados.

Os objetivos da análise de redes resumiram-se a analisar a relação entre laboratórios (de produção dos medicamentos) e tarja, a relação entre classes terapêuticas, faixa de preço de revenda e a relação entre laboratórios e faixa de preço de fábrica, classe terapêutica e regime de preço via projeção bipartida com geração de uma rede monopartida de classe terapêutica.

As redes brutas geradas foram organizadas utilizando os métodos de distribuição Yifan Hu e Fruchterman Reingold, cujos algoritmos estão disponíveis no software Gephi. A visualização dos dados utilizando *dashboard* foi desenvolvida no software Microsoft Power BI Desktop[7] a partir da base de dados em extensão '.xlsx'

[2] O software OpenRefine é uma ferramenta de código aberto utilizada para a limpeza e transformação de dados. Disponível em https://openrefine.org/.

[3] O formato camelCase integra o uso de vários softwares e linguagens por meio da padronização sugerida na terminologia de nomes de variáveis. Disponível em: https://pt.wikipedia.org/wiki/CamelCase.

[4] *General Refine Expression Language* (GREL). É uma linguagem de script similar ao javascript utilizada no ambiente do OpenRefine.

[5] GML (Graph Modelling Language) é um formato para representação de grafos de fácil leitura por humanos e com uma capacidade semântica razoável para configurar as características da rede, dos nós e das arestas. Disponível em: https://en.wikipedia.org/wiki/Graph_Modelling_Language/.

[6] GEPHI é um software de código aberto utilizado para visualização, análise e manipulação de redes e grafos. Disponível em https://gephi.org/.

[7] O Microsoft Power BI Desktop é um conjunto de serviços de software de uso gratuito destinados à manipulação de dados com o propósito de gerar, por exemplo, painéis interativos. Disponível em https://powerbi.microsoft.com/.

exportada pela ferramenta OpenRefine, após as etapas de pré-processamento e transformação. O *dashboard* foi construído a partir da seleção de laboratórios, tarja e regime de preço, exibindo a comparação entre preço de fábrica e preço de revenda para cada produto, a distribuição de todos produtos em faixas de preço de fábrica e o número de medicamentos por taja.

4 Resultados

Na etapa de mineração destacaram-se algumas redes envolvendo os seguintes nós: classes de preço de fábrica, laboratórios, e tarja. O relacionamento entre as variáveis tarja e laboratórios, Figura 3, gerou uma rede que sugere uma predileção dos laboratórios pela produção de medicamentos de tarja vermelha, que são vendidos apenas sob a prescrição de médicos ou dentistas. Entre os principais medicamentos de tarja vermelha comercializados no Brasil estão os fármacos para o tratamento de diabetes, hipertensão e medicamentos psicotrópicos.

Na mesma rede nota-se um alto número de registros onde a tarja do produto não é informada, além de observar 11 laboratórios que não informaram a tarja de nenhum dos seus medicamentos produzidos. Além disso, nenhum dos laboratórios produz apenas medicamentos de tarja preta, diferente do que ocorre com as tarjas vermelha e venda livre.

Quando relacionamos as variáveis faixas de preço de fábrica e laboratórios, Figura 4, observou-se que a maioria dos laboratórios tem em seu catálogo produtos cujo preço para os varejistas custa menos do que R$1000, sendo a faixa *Menos que R$100* a maior parcela entre todas. O número de laboratórios vai diminuindo conforme aumenta-se o valor dos produtos, chegando ao extremo onde apenas um laboratório produz medicamentos com valor de fábrica acima de 1 milhão de reais. Nesta mesma rede, também foi possível observar uma significativa quantidade de laboratórios que produzem apenas medicamentos com valor de fábrica abaixo de R$100. Ainda, dois dos 268 laboratórios produzem apenas medicamentos com custo acima de R$50000 e abaixo de 1 milhão de reais.

A rede monopartida, Figura 5, formada por nós da Classe Terapêutica por meio de projeção bipartida com as variáveis Faixa de Preço de Fábrica e Regime de Preço possui 534 nós, equivalentes às diferentes classes terapêuticas. Ao realizar o cálculo de modularidade, foram gerados 5 agrupamentos de classes terapêuticas identificados na rede por cores:

- Marrom: 224 nós
- Azul: 115 nós
- Verde: 97 nós
- Roxo: 40 nós
- Amarelo: 36 nós
- Vermelho: 22 nós

Destaca-se, por exemplo, que o grupo verde é composto por 60% de classes terapêuticas de venda livre (tarja), enquanto que o segundo grupo com maior proporção desta categoria de tarja conta com apenas 10%. Com relação ao tipo de produto, o agrupamento verde também é o que possui o maior número de fitoterápicos, com 8,32%, enquanto que em todos os outros grupos esta porcentagem não chega a 1%. Ainda, no grupo marrom há aproximadamente o dobro ou mais de classes terapêuticas com restrição hospitalar quando comparado aos outros agrupamentos.

O *dashboard* gerado com o software Power BI Desktop, Figura 6, permitiu observações amplas da base de dados como, por exemplo, o número total de laboratórios, produtos e substâncias e o cálculo, por exemplo, do valor médio do preço de fábrica de um subconjunto específico de dados. O painel possibilitou a segmentação precisa da base de dados, como a escolha de um ou vários laboratórios e o filtro de tarja e de regime de preço. Ainda, foi possível visualizar em gráfico de barras a comparação entre o preço de fábrica e o preço de revenda dos produtos, a distribuição de produtos por faixa de preço e a quantidade de produtos em cada categoria de tarja. No total, observou-se que a base de dados com 268 laboratórios tem registro de 6242 produtos e 2307 substâncias, sendo que a média do preço de fábrica entre todos os produtos é de aproximadamente R$3000. Nota-se também que a proporção de produtos entre as faixas de preço se manteve independentemente da tarja.

5 Considerações Finais

Dentre as categorias de tarja, a tarja vermelha é a que é produzida pelo maior número de laboratórios, o que pode se relacionar com um cenário em que a hipertensão arterial é uma das principais causas de morte no Brasil, que acomete 24% das pessoas com mais de 18 anos, e em que a diabetes ocorre em 7% da população. Ambas as condições clínicas são tratadas por remédios controlados de tarja vermelha, em que é obrigatória a prescrição médica. Outras condições de saúde com diferentes prevalências também podem contribuir para o aumento da demanda de medicamentos de tarja vermelha.

Muitos registros de produtos não incluíram a informação de tarja, o que prejudica a análise assertiva da base de dados. Imagina-se que a não inclusão desse tipo de informação é devida à falhas humanas no processo de alimentação da base de dados. Essa situação ilustra um problema grave da gestão de dados e, se tratando de dados públicos, mantidos por uma instituição pública e relacionados com uma temática tão cara à população, calcula-se que o prejuízo gerado pela má gestão dos dados seja potencialmente alto.

Alguns laboratórios se reservam a produzir apenas um tipo específico de tarja, embora nenhum dos 268 laboratórios produza exclusivamente medicamentos de tarja preta. Os medicamentos de tarja preta são os psicoativos que têm um alto potencial de causar dependência, como as morfinas e anfetaminas e, por isso, deduz-se que seu uso restrito limite a demanda destes no mercado, talvez impossibilitando um modelo de negócios farmacêutico com produção exclusiva.

A maioria das classes terapêuticas registradas na base de dados tem o preço de fábrica girando em torno de R$100, enquanto que apenas duas classes terapêuticas tem custo acima de 1 milhão de reais. São elas *M5X - TODOS OS OUTROS FÁRMACOS COM AÇÃO MÚSCULO-ESQUELÉTICA* e *S1X1 - OUTROS PRODUTOS OFTALMOLÓGICOS SISTÊMICOS.* Ambas as classes terapêuticas são produzidas por um único laboratório, Novartis Biociências S.A., que produz todos os produtos com valor acima de um milhão de reais.

As relações intracluster da rede monopartida ainda precisam ser investigadas de maneira mais minuciosa a fim de entender de maneira plena quais as características fundamentais de cada um dos agrupamentos e a sua relevância para esta pesquisa

Portanto, o objetivo da pesquisa foi parcialmente alcançado uma vez que algumas relações entre as variáveis foram descobertas e investigadas. Contudo, é necessária a continuação do esforço de trabalho para analisar possíveis relações ainda não descobertas na base de dados manipulada neste estudo.

Referências

ALBERT, Réka; BARABÁSI, Albert-László. Statistical mechanics of complex networks. **Reviews of modern physics**, v. 74, n. 1, p. 47, 2002.

BARABÁSI, Albert-László. **Linked: The new science of networks**. 2003.

BÁRRIOS, Maria João; MARQUES, Rita; FERNANDES, Ana Alexandre. Aging with health: aging in place strategies of a Portuguese population aged 65 years or older. **Revista de Saúde Pública**, v. 54, 2020.

BATISTA, Gustavo Enrique de Almeida Prado et al. **Pré-processamento de dados em aprendizado de máquina supervisionado**. 2003. Tese de Doutorado. Universidade de São Paulo.

COSTA, Claudio Napolis et al. Descoberta de conhecimento em bases de dados. **Revista Eletrônica: Faculdade Santos Dumont**, v. 2, p. 20, 2019.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v. 39, n. 11, p. 27-34, 1996.

HAND, David J. Principles of data mining. **Drug safety**, v. 30, n. 7, p. 621-622, 2007.

METZ, Jean et al. Redes complexas: conceitos e aplicações. 2007.

WASSERMAN, Stanley et al. Social network analysis: Methods and applications. 1994.

Apêndice A

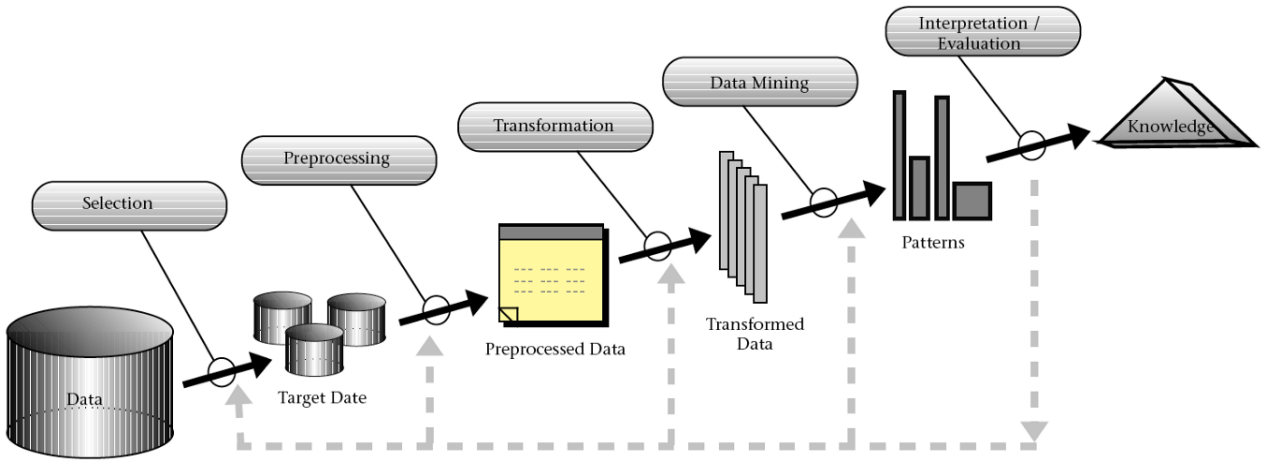


Figura 1 - Etapas do processo KDD.
Fonte: Fayyad et al. (1996).

```
graph [
directed 0


node [ id {{jsonize(cells.classeTerapeutica.value)}}
variavel "Classe Terapeutica" agrupamento "grupo A" ]

node [ id {{jsonize(cells.faixaPrecoFabrica17.value)}}
variavel "Faixa de Preço de Fabrica" agrupamento "grupo
B" ]

node [ id {{jsonize(cells.regimeDePreco.value)}}
variavel "Regime de Preço" agrupamento "grupo B" ]


edge [ source {{jsonize(cells.classeTerapeutica.value)}}
target {{jsonize(cells.faixaPrecoFabrica17.value)}}]

edge [ source {{jsonize(cells.classeTerapeutica.value)}}
target {{jsonize(cells.regimeDePreco.value)}}]

]
```

Figura 2 - Código de mapeamento GML pelo software OpenRefine para geração de rede tripartida.
Fonte: autoria própria.

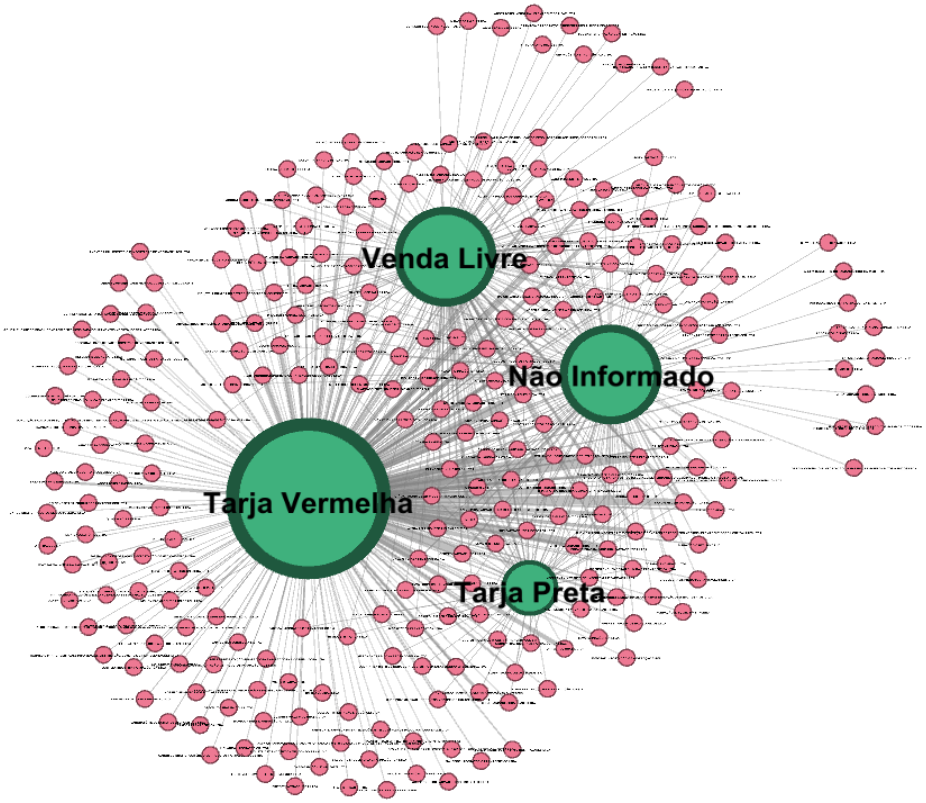


Figura 3 - Rede informacional bipartida relacionando as variáveis tarja e laboratório.
Fonte: autoria própria, com apoio do software Gephi.

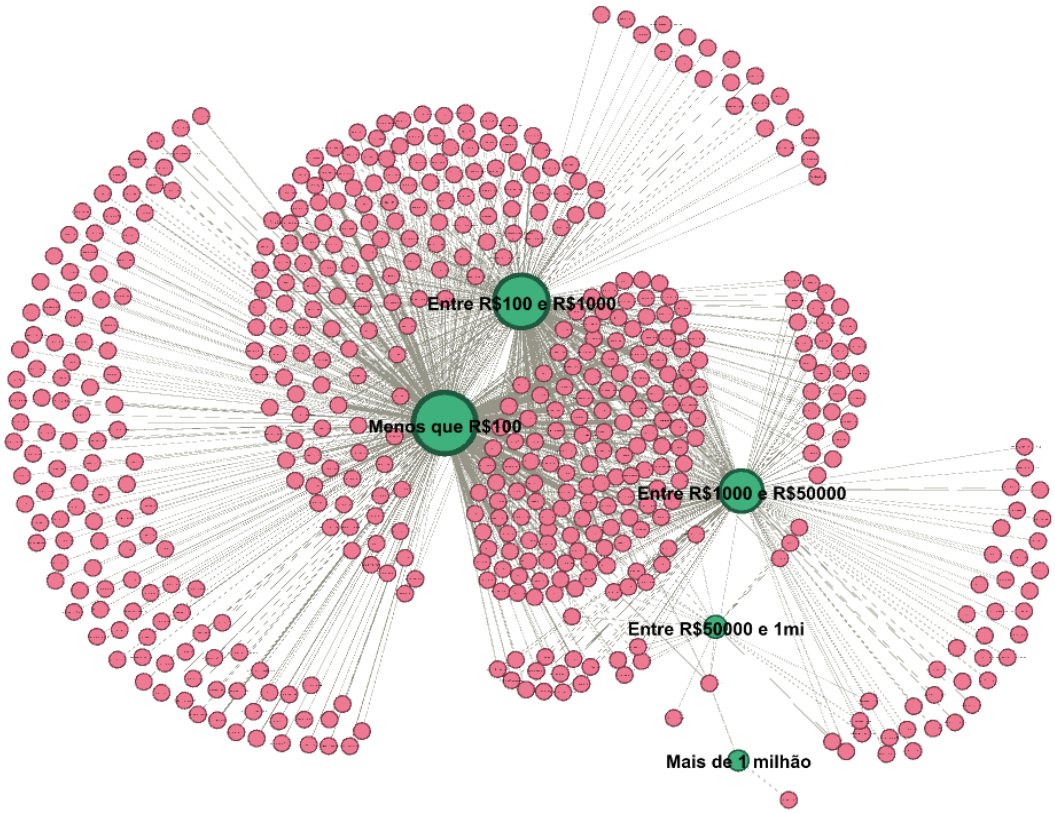


Figura 4 - Rede informacional bipartida relacionando as variáveis faixa de preço de fábrica e classe terapêutica.
Fonte: autoria própria, com apoio do software Gephi.

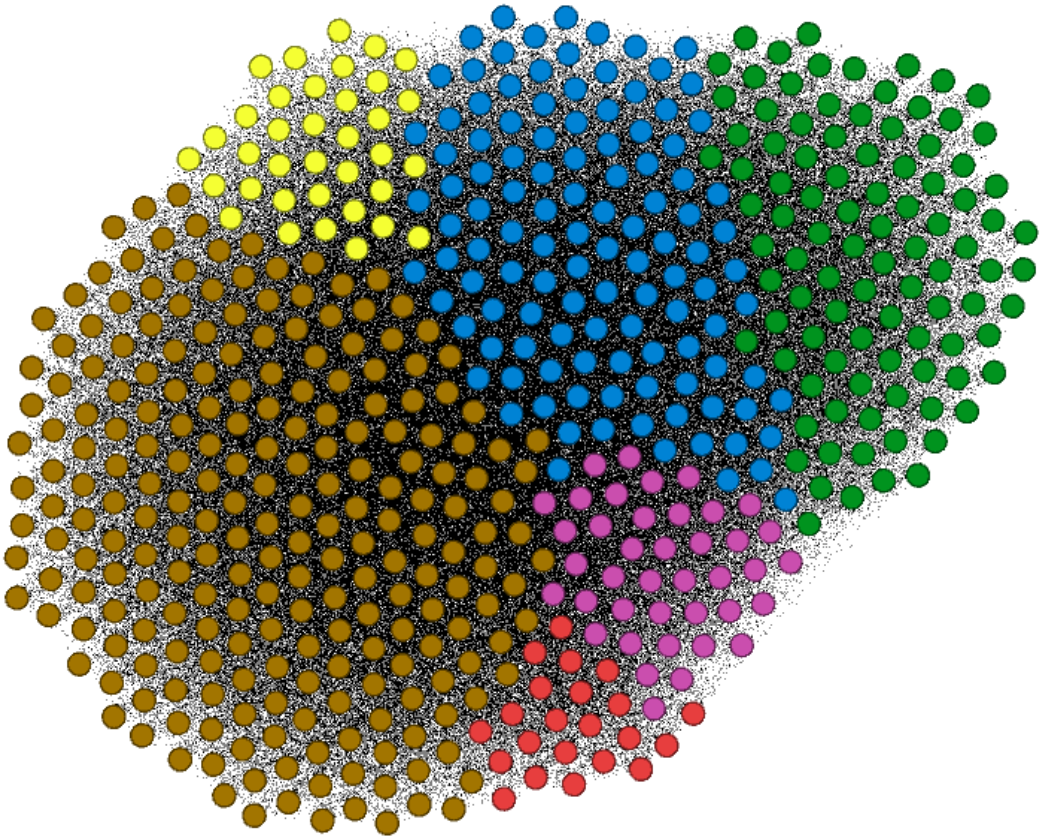


Figura 5 - Rede informacional monopartida relacionando as variáveis faixa de preço de fábrica, classe terapêutica e regime de preço colorida após cálculo de modularidade.
Fonte: autoria própria, com apoio do software Gephi.

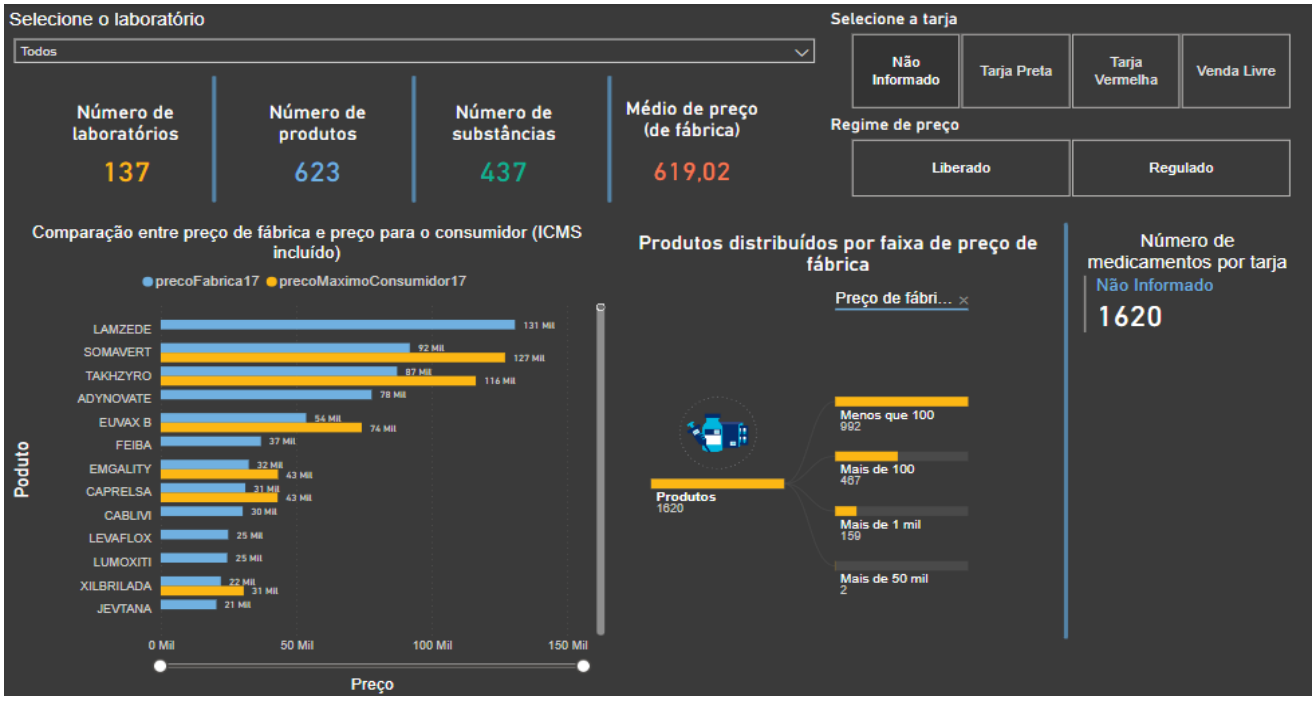


Figura 6 - Dashboard de variáveis da base de dados
Fonte: autoria própria, com apoio do software Power BI Desktop

DESCOBERTA DE RELAÇÕES ENTRE ESTADOS BRASILEIROS A PARTIR DE DADOS FINANCEIROS DE OPERAÇÕES DE CRÉDITO DISPONÍVEIS EM DADOS ABERTOS DO BANCO CENTRAL

DISCOVERY OF RELATIONS BETWEEN BRAZILIAN STATES FROM FINANCIAL DATA OF CREDIT OPERATIONS AVAILABLE IN OPEN DATA OF The Central Bank of Brasil

Juliana Rodrigues de Lima Meirelles[1], Henrique Monteiro Cristovão[2], Daniela Lucas da Silva Lemos[3]

(1) Universidade Federal do Espírito Santo, Vitória-ES, juliana.meirelles@edu.ufes.br
(2) Universidade Federal do Espírito Santo, Vitória-ES, henrique.cristovao@ufes.br
(3) Universidade Federal do Espírito Santo, Vitória-ES, daniela.l.silva@ufes.br

Resumo

No sistema bancário brasileiro mudanças têm impactado a forma como a sociedade lida com suas finanças, e cada vez mais o uso dos dados é enfatizado como sendo o fator de competitividade de um produto para o outro. O objetivo geral é investigar e evidenciar relacionamentos existentes entre os estados brasileiros sobre aspectos do consumo de produtos de crédito bancário. Para tanto, foram utilizados dados financeiros e de operações de crédito a partir de dados abertos do Banco Central. Consiste em pesquisa aplicada, de caráter exploratório e descritivo, que visa não só relacionar as variáveis de análise central, mas também apresentar informações que possam servir de diretrizes para ações de transformação da realidade. Foi feita uma modelagem conceitual sobre um grupo de variáveis que representassem o problema de pesquisa. Utilizou-se técnicas de análise de redes de informação para revelar relacionamentos não aparentes entre as variáveis selecionadas. Condensando os resultados, pôde-se observar os estados de SP, RS, SC, PR e MG como sendo relevantes perante à sociedade, nos aspectos relativos ao consumo de produtos bancários, bem como o incentivo econômico gerado pelos mesmos, alavancando também o consumo de outros estados dependentes. O desenvolvimento deste estudo, bem como o exame da literatura mostrou-nos que o cenário atual encontra-se promissor para aplicação de técnicas de análise de redes de informação, pois possibilita gerar insights com execuções interessantes principalmente para a área de gestão da informação e do conhecimento.

Palavras-chave: análise de redes de informação; dados abertos; rede monopartida; modelagem de dados; mercado financeiro.

Abstract

In the Brazilian banking/financial system, changes have impacted the way society deals with its finances and the use of data is increasingly emphasized as a factor of competitiveness from one product to another (VALENTIM, 2006). The present study sets out a research problem: What are the most relevant relationships between Brazilian states on aspects of the consumption of bank credit products? Thus, the general objective becomes to investigate and bring relevant relationships between Brazilian states on aspects of the consumption of bank credit products. For this purpose, financial data and credit operations based on open data from the Central Bank will be used.It consists of applied research, of an exploratory and descriptive nature, which aims not only to relate the variables of central analysis, but also to present information that can serve as guidelines for actions to transform reality. Condensing the results, one can observe the states of SP, RS, SC, PR and MG as being relevant to society, in the aspects related to the consumption of banking products, as well as the economic incentive generated by them, also leveraging the consumption of other dependent states. The development of this study, as well as the examination of the literature, showed us that the current scenario is promising for network science. With the study, we were able to observe that the practical applicability of this technique in the financial area is satisfactory and promising, as it makes it possible to generate insights with interesting executions, mainly for the area of information and knowledge management.

Keywords: analysis of information networks; open data; one-party network; data modeling; financial market.

1 Introdução

O mundo passa por grandes transformações no que tange à gestão e uso da informação. No sistema bancário/financeiro brasileiro não é diferente, pois mudanças têm impactado a forma como a sociedade lida com suas finanças e cada vez mais o uso dos dados é enfatizado como

sendo o fator de competitividade de um produto para o outro (VALENTIM, 2006). Gonzalez (2016) argumenta que a demanda por serviços financeiros mais eficientes e apropriados ainda é grande. Isso explica a crescente criação de *startups*, considerado por Gitahy (2011) como um grupo de pessoas à procura de um modelo de negócios repetível e escalável. Inclui-se aí as *FinTechs* que são orientadas pela tecnologia, não estando vinculados a sistemas legados (JUENGERKES, 2016). Outro fator importante e atual do setor financeiro é o surgimento do *Open Banking*, ou sistema financeiro aberto. Ele promete revolucionar a forma como se lida com produtos, serviços e dados financeiros dos cidadãos. Em uma primeira fase do Open Banking, houve o compartilhamento dos dados das instituições financeiras de forma padrão, pelo Banco Central. Essa nova realidade traz consigo a inovação, o surgimento de novos modelos de negócios, difusão de informação, favorecendo a inclusão e a educação financeira da população, possibilitando, por exemplo, o surgimento de aplicativos de comparação de preços de produtos financeiros (OPEN BANKING BRASIL, 2022).

Segundo Valentim (2008), deve-se refletir sobre os dispositivos que transformam dados em informação. Saber lidar com o alto volume, diferentes origens e formatos de dados é uma das habilidades mais valorizadas das últimas décadas (DAVENPORT et al., 2012 apud ISOTANI, 2015). A utilização da modelagem conceitual de dados é um conhecimento já consolidado no mercado de trabalho, principalmente para o desenvolvimento de softwares, e na Ciência da Informação encontra-se presente na organização e representação da informação como forma de fornecer uma descrição estável e coerente dos dados (SAYÃO, 2001). Segundo Guizzardi (2005), a modelagem está associada à representação das coisas do mundo, independente de escolhas tecnológicas, e ela ainda objetiva entender, comunicar e raciocinar sobre uma porção da realidade. No mesmo intuito, a visualização da informação visa aumentar a compreensão do usuário sobre algum conhecimento, trazendo assim a inclusão informacional dos usuários (DIAS, 2007).

Nossa sociedade está cada vez mais dependente de redes e se pudéssemos resumir em uma palavra a atual sociedade esta palavra seria “conectada” (CAVALCANTE, 2009) e ainda, de acordo com Barabási (2013), várias redes surgem e evoluem baseados em leis e mecanismos que são a base para a nova ciência chamada de Ciência das Redes.

A análise de redes sociais preocupa-se com o entendimento das ligações entre entidades sociais e as implicações destas ligações (WASSERMAN; FAUST, 1994). Para a análise de redes de informação, também conhecida por análise de redes complexas, utilizam-se praticamente as mesmas técnicas da análise de redes sociais. Considerada como um dos métodos de apoio à descoberta de conhecimento na Ciência de Dados, a análise de redes de informação é uma prática que tem ajudado muitos pesquisadores a descobrirem relacionamentos não aparentes entre variáveis de uma base de dados.

Uma rede denominada de monopartida é composta por nós advindos de apenas uma variável e é útil para revelar relacionamentos existentes entre os grupos de nós sobre essa variável. Normalmente ela é obtida de projeções bipartidas, que é um processo de eliminação de variáveis de uma rede.

O Banco Central do Brasil divulga mensalmente informações agregadas das operações de crédito realizadas no âmbito nacional pelas instituições autorizadas, recebidas através do Sistema de Informações de Créditos – SCR. São disponibilizados dados agregados de Carteira Ativa, Inadimplência e Ativo Problemático com possibilidade de detalhamento por tipo de cliente (PF/PJ), modalidade de crédito, unidade da federação, natureza da ocupação (PF), porte/rendimento dos clientes, origem de recursos e indexador das operações (Painel de Operações de Crédito, 2022). Esses dados são disponibilizados no Portal de Dados Abertos do Banco Central onde é possível fazer os downloads das séries em diversos formatos.

Questões relativas à gestão da informação, como acessá-las e difundi-las, tornaram-se áreas discutidas por estudiosos da Ciência da Informação, pois é sabido que

os sistemas de informação, por si só, não são suficientes para responder às demandas informacionais das pessoas e organizações. Tendo em vista a relevância dos impactos das mudanças no sistema bancário/ financeiro brasileiro para as sociedades, e a necessidade da incorporação definitiva do tema nas agendas dos governos locais em todas as partes do mundo, o presente estudo estabelece como problema de pesquisa: existem relacionamentos relevantes entre os estados brasileiros sobre aspectos do consumo de produtos de crédito bancário?

Dessa forma, há necessidade de um direcionamento da oferta no setor de crédito, que poderá ser utilizado socialmente em forma de incentivos direcionados pelo governo a nichos específicos de consumidores de produtos de crédito, como forma de alavancar o consumo e consequentemente a economia. Segundo Matta (2010), em se tratando de aspectos informacionais, é fato que todos os envolvidos em um processo de mudança de comportamento necessitam, buscam e utilizam informações para alcançar os objetivos almejados.

2 Objetivos

O objetivo geral é investigar e evidenciar os relacionamentos existentes entre os estados brasileiros sobre aspectos do consumo de produtos de crédito bancário. Para tanto, serão utilizados dados financeiros e de operações de crédito a partir de dados abertos do Banco Central[1].

3 Procedimentos Metodológicos

O presente estudo consiste em uma pesquisa de natureza aplicada, de caráter exploratório e descritivo, que visa não só relacionar as variáveis de análise central, mas também apresentar informações que possam servir de diretrizes para ações de transformação da realidade.

Os resultados são apresentados sob a forma qualitativa, a partir da coleta de informações de fontes primárias e secundárias, incluindo revisão bibliográfica. A planificação da pesquisa inclui o levantamento dos dados e a revisão da literatura. A análise está direcionada ao contexto que configura o objeto de estudo, a saber, o sistema bancário brasileiro. Para a pesquisa bibliográfica foram utilizados autores relevantes para a Ciência da Informação. A coleta de dados foi realizada utilizando dados financeiros e de operações de crédito a partir de dados abertos do Banco Central.

Foi utilizada a base de dados denominada “SCR.data - Painel de Operações de Créditos”. O Banco Central do Brasil divulga mensalmente informações agregadas das operações de crédito recebidas através do Sistema de Informações de Créditos – SCR. São disponibilizados dados agregados de Carteira Ativa, Inadimplência e Ativo Problemático com possibilidade de detalhamento por tipo de cliente (PF/PJ), modalidade de crédito, unidade da federação, Classificação Nacional de Atividades Econômicas – CNAE (PJ), natureza da ocupação (PF), porte/rendimento dos clientes, origem de recursos e indexador das operações. A publicação compreende o período de junho de 2012 até a atualidade e é baseada no documento 3040 (SCR), com informações detalhadas de todas as operações de crédito cursadas no país de valor superior a R$ 1.000 até a data-base de maio/16 e de valor superior a R$ 200 a partir da data-base junho/16. O total nacional mais 27 unidades da federação estão presentes no relatório.

A presente pesquisa usou 840.114 registros de crédito, que correspondem ao mês de maio de 2022. Das 22 variáveis (campos) existentes na base de dados, foram selecionadas sete para viabilizar a realização da pesquisa, para um primeiro momento. Nessa seleção das variáveis usou-se o critério daquelas que mais se aproximavam do problema de pesquisa. Para tanto foi consultado o dicionário de dados denominado ”metodologia” disponível no site.

A modelagem conceitual foi construída sobre as variáveis selecionadas, usando-se um modelo de entidade e relacionamento

[1] Portal de dados abertos do Banco Central do Brasil, disponível em: Disponível em: https://dadosabertos.bcb.gov.br/dataset/scr_data. Acesso em: 12 ago. 2022.

(MER)[2] e a ferramenta de diagramação DBDesigner[3]. Foi realizada identificando duas entidades centrais, a saber, Clientes e Créditos. Seguindo os princípios de normalização dos dados, identificamos outras entidades relacionadas aos Clientes: UF, Ocupação, e Porte; e aos Créditos: Indexador e Modalidade.

A ferramenta Open Refine[4] foi usada para limpeza e preparação de dados, realizando a exportação dos dados para o formato GML[5] para, posteriormente, serem abertos na ferramenta Gephi[6].

Finalmente, utilizando-se de técnicas de análise de redes complexas, formatações adequadas e projeções bipartidas sobre as redes obtidas, pôde-se revelar os relacionamentos que são apresentados na seção de resultados.

4 Resultados

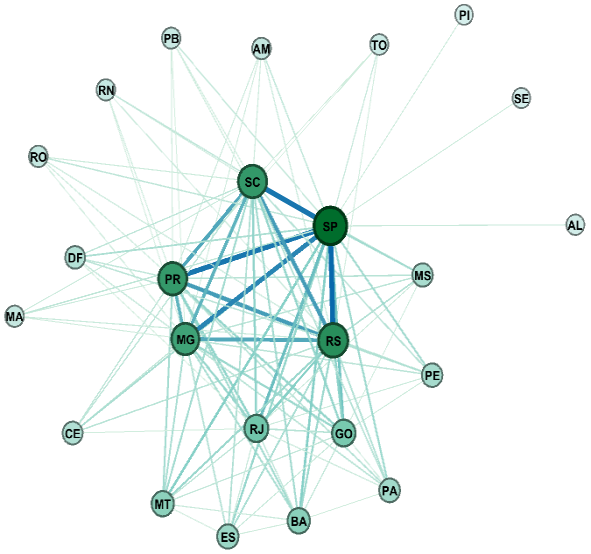
Na modelagem conceitual, sobre as sete variáveis escolhidas da base de dados, Figura 4 do Apêndice A, foram identificadas as principais entidades do modelo: Cliente e Crédito. Em seguida identificou-se os atributos de cada uma das entidades, a normalização dos dados, e os relacionamentos entre os atributos dos consumidores (Cliente) e do Crédito.

Após a preparação dos dados e mapeamento em formato de rede GML, foram construídas três redes monopartidas de estados brasileiros para realização de análise comparativa pela modalidade de crédito, pelo porte salarial e pela ocupação.

A Figura 1 apresenta a rede monopartida de estados obtida por uma projeção bipartida pela eliminação da variável modalidade de crédito. As modalidades das operações de crédito são agrupadas em:

• PF - Cartão de crédito
• PF - Empréstimo com consignação em folha • PF - Empréstimo sem consignação em folha • PF - Habitacional
• PF - Outros créditos
• PF - Rural e agroindustrial
• PF - Veículos

Figura 1 - Rede monopartida de estados, por modalidade de crédito



Fonte: autoria própria, com apoio do software Gephi

Na inspeção visual da rede da Figura 1, os estados que mais se destacaram no relacionamento entre as modalidades de crédito foram SP, RS, SC, PR e MG. Isso demonstra que eles se assemelham em volume, nas modalidades de crédito mais contratadas pelos brasileiros. SP, como esperado, tem um relacionamento mais forte com outros estados. O RS aparece como sendo o segundo lugar em contratações de crédito no Brasil. Podemos ver também RJ, e GO em segundo plano, porém fortes em relação ao volume de contratações dos outros estados.

A Figura 2 apresenta a rede monopartida de estados obtida por uma projeção bipartida

[2] MER é um modelo de dados para descrever os dados ou aspectos de informação de um domínio de negócio ou seus requisitos de processo. Disponível em: https://pt.wikipedia.org/wiki/Modelo_entidade_relacionamento.

[3] A DBDesigner é uma ferramenta CASE para a modelagem de dados que trabalha com o modelo lógico. Disponível em: https://www.devmedia.com.br/dbdesigner-modelagem-de-dados.

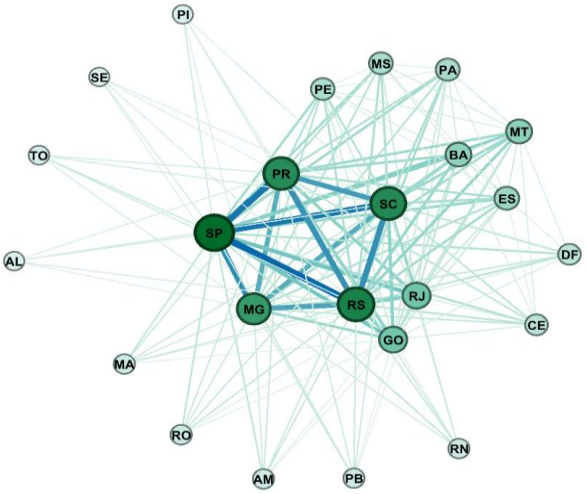
[4] OpenRefine é um software para limpeza, preparação e reconciliação de dados. Disponível em: https://openrefine.org/https://www.devmedia.com.br/dbdesigner-modelagem-de-dados.

[5] GML (*graph modelling language*) é um formato universal para representação de grafos. Disponível em: https://en.wikipedia.org/wiki/Graph_Modelling_Language.

[6] Gephi é um software para visualização, exploração e análise de vários tipos de redes e sistemas complexos, grafos dinâmicos e hierárquicos. Disponível em: https://gephi.org/.

pela eliminação da variável porte salarial. As opções de porte salarial dos clientes são:
• PF - Até 1 salário-mínimo
• PF - Mais de 1 a 2 salários-mínimos
• PF - Mais de 2 a 3 salários-mínimos
• PF - Mais de 3 a 5 salários-mínimos
• PF - Mais de 5 a 10 salários-mínimos
• PF - Mais de 10 a 20 salários-mínimos
• PF - Acima de 20 salários-mínimos

Figura 2 - rede monopartida de estados, por porte salarial

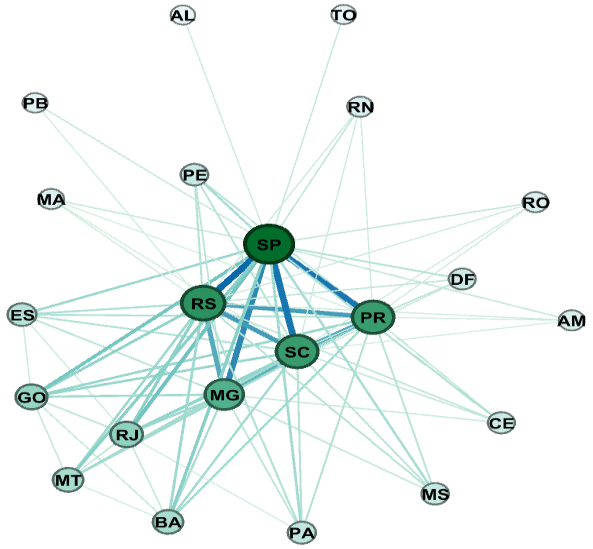


Fonte: autoria própria, com apoio do software Gephi

Semelhantemente, na Figura 2, os estados que mais se destacaram no relacionamento entre os estados em função do porte dos clientes foram SP, RS, SC, PR e MG. Isso demonstra que eles se assemelham às características de níveis salariais. SP, aparece também em primeiro lugar, seguido de forma muito próxima pelo RS. No entanto, a força dos relacionamentos entre os estados aparece de forma preponderante. Pode-se ver novamente RJ, e GO em segundo plano, porém agora em relação ao porte dos consumidores, BA, MT e ES evidenciam também relações coesas com os estados com maior volume.

A Figura 3 apresenta a rede monopartida de estados obtida por uma projeção bipartida pela eliminação da variável ocupação. As naturezas de ocupação referem-se às pessoas físicas e estão agrupadas da seguinte forma:
• PF - Servidor ou empregado público
• PF - Empregado de entidades sem fins lucrativos
• PF - Empregado de empresa privada
• PF - Aposentado/pensionista
• PF - Autônomo
• PF - Empresário
• PF - MEI

Figura 3 - rede monopartida de estados, por ocupação



Fonte: autoria própria, com apoio do software Gephi

Os estados que mais se destacaram, na Figura 3, em função do relacionamento com a natureza de ocupação das pessoas físicas foram SP, seguidos dos estados RS, SC, PR e MG. Os estados do RS, SC E MG aparecem com volume de dados semelhantes. SP, mais uma vez, como esperado, tem um relacionamento mais forte com outros estados. Podemos ver também RJ, GO, MT e BA em segundo plano, porém com volumes semelhantes e mais fortes em relação aos outros estados.

Condensando os resultados, cabe portanto salientar o quanto os estados SP, RS, SC, PR e MG são relevantes perante à sociedade, nos aspectos relativos ao consumo de produtos bancários, bem como o incentivo econômico gerado pelos mesmos, alavancando também o consumo de outros estados mais dependentes.

Os grafos obtidos através da análise dos dados abertos do banco central possuem grande potencial para abstrair informações relevantes para a área financeira. Segundo Passos (2020) a teoria de redes complexas

pode ser utilizada ainda para descrever e modelar sistemas de direcionamento de produtos e serviços das empresas, criando um sistema de recomendação. Com a projeção de rede bipartida é possível comparar o método com outras formas de recomendação. São inúmeras as possibilidades diante das bases abertas do Banco Central, selecionando por exemplo as tags relativas às tarifas teria-se um arcabouço para a criação de um comparador de instituições financeiras em diversos aspectos que trariam informações relevantes do ponto de vista da sociedade e sem conflitos de interesses devido à natureza dos dados utilizados.

4 Considerações Finais

Como a Ciência das Redes se configura como área relativamente nova no âmbito das ciências, as pesquisas estão se ajustando e se desenvolvendo em conjunto com novas tecnologias. A exploração de sua interdisciplinaridade, a análise da aplicabilidade de ferramentas, e a descoberta de relações com outros saberes e informações, emerge, então, como uma área promissora para pesquisa e para a Ciência da Informação.

O desenvolvimento deste estudo, bem como o exame da literatura mostrou que o cenário atual encontra-se promissor para aplicação das técnicas empregadas. Com o estudo pode-se observar que a aplicabilidade prática dessa técnica na área financeira é satisfatória e promissora, pois possibilita gerar insights com execuções interessantes principalmente para a área de gestão da informação e do conhecimento.

A partir dos dados abertos sobre operações de crédito, disponibilizados pelo Banco Central, foi realizada uma modelagem conceitual e a construção de três redes monopartidas de estados considerando-se a modalidade de crédito, o porte salarial e a ocupação. Contudo, utilizou-se um número reduzido de variáveis como forma de viabilizar o experimento.

Nas três redes monopartidas destacaram-se os estados SP, RS, SC, PR e MG, observando-se que eles são relevantes perante à sociedade, nos aspectos relativos ao consumo de produtos bancários.

O objetivo proposto, a saber, de investigar, e trazer relacionamentos relevantes entre os estados brasileiros sobre aspectos do consumo de produtos de crédito bancário foi alcançado e evidenciado com a modelagem conceitual e, principalmente, pela análise das redes geradas por meio de inspeção visual.

Os resultados da pesquisa podem orientar a aplicação de incentivos direcionados a nichos específicos de consumidores de produtos de crédito, para incentivar o consumo em determinada área ou estado, e consequentemente estimular a economia.

A inspeção visual s redes demonstraram que existem relacionamentos relevantes entre os estados brasileiros sobre aspectos do consumo de produtos de crédito bancário, e que podem ser ainda explorados sob diversos aspectos da base de dados utilizada, visando trabalhos futuros com a mesma base ou outras subjacentes que venham tratar com questões interessantes sobre por exemplo a inadimplência da população.

Acredita-se que este trabalho seja um passo inicial para contribuir com os diversos usuários do sistema financeiro, mais especificamente do mercado de crédito bancário, no sentido de auxiliar sobre o direcionamento de investimentos a partir da constatação de que existem relacionamentos relevantes entre os estados brasileiros no que tange a produtos de crédito bancário, porte salarial, e natureza da ocupação dos usuários.

Nesse sentido, um exemplo de aplicação poderia ser a partir de órgãos públicos estaduais ou governamentais que tenham interesse no direcionamento de créditos de incentivo financeiro para a população, incentivo fiscal para a indústria e comércio da região que tenham características comuns, ou mesmo políticas de educação financeira da população. Podem ser determinadas políticas de incentivo ao consumo de determinado produto de crédito como redução de juros à medida que fique evidenciada a necessidade latente da população, como exemplo, os créditos rural, agro e de habitação de dada região. Podem ser verificados também perfis comportamentais de consumo dos

habitantes, em função da faixa salarial e de natureza da ocupação de determinado estado ou região que aparecem nos grafos de forma semelhante.

Referências

BARABÁSI, Albert-László. Ciência da rede. Transações Filosóficas da Royal Society A: Mathematical, Physical and Engineering Sciences, v. 371, n. 1987, p. 20120375, 2013.

CAVALCANTE, Gustavo Vasconcellos. Ciência das Redes: aspectos epistemológicos. 2009.

DIAS, Mateus Pereira; CARVALHO, José Oscar Fontanini de. A Visualização da Informação e a sua contribuição para a Ciência da Informação. DataGramaZero, Rio de Janeiro, v. 8, n. 5, p. 1-16, 2007.

GITAHY, Yuri. O que é uma startup. Empreendedor Online–Empreendedorismo na Internet e negócios online, 2011.

GONZALEZ, L.; CERNEV, A. K.; DINIZ, E. O. Desafio da Inclusão Financeira e a Promessa das Fintechs. Sistema Financeiro Nacional: o que fazer, 2016

GUIZZARDI, Giancarlo. Fundamentos ontológicos para modelos conceituais estruturais. 2005.

ISOTANI, Seiji; BITTENCOURT, Ig Ibert. Dados abertos conectados: em busca da web do conhecimento. Novatec Editora, 2015.

JUENGERKES, Bjoern Erik. FinTechs and Banks–Collaboration is Key. The FinTech Book: The Financial Technology Handbook for Investors, Entrepreneurs and Visionaries, p. 179-182, 2016

Mais liberdade de escolha e serviços para você. Open Banking Brasil, 2022. Disponível em: https://openbankingbrasil.org.br/?cookie=true. Acesso em: 18 set. 2022.

MATTA, Rodrigo Octávio Beton. Modelo de comportamento informacional de usuários: uma abordagem teórica. Gestão, mediação e uso da informação, p. 127-142, 2010.

Painel de Operações de Créditos. 22 jul. 22. Disponível em: https://dadosabertos.bcb.gov.br/dataset/scr_data. Acesso em: 12 ago. 2022.

PASSOS, Gabriela. Redes Bipartidas para recomendação de produtos financeiros. 2020. Tese de Doutorado. Universidade de São Paulo.

VALENTIM, Marta Lígia Pomim. Capítulo 1 PROCESSO DE INTELIGÊNCIA COMPETITIVA ORGANIZACIONAL. Informação, conhecimento e inteligência organizacional, p. 9, 2006.

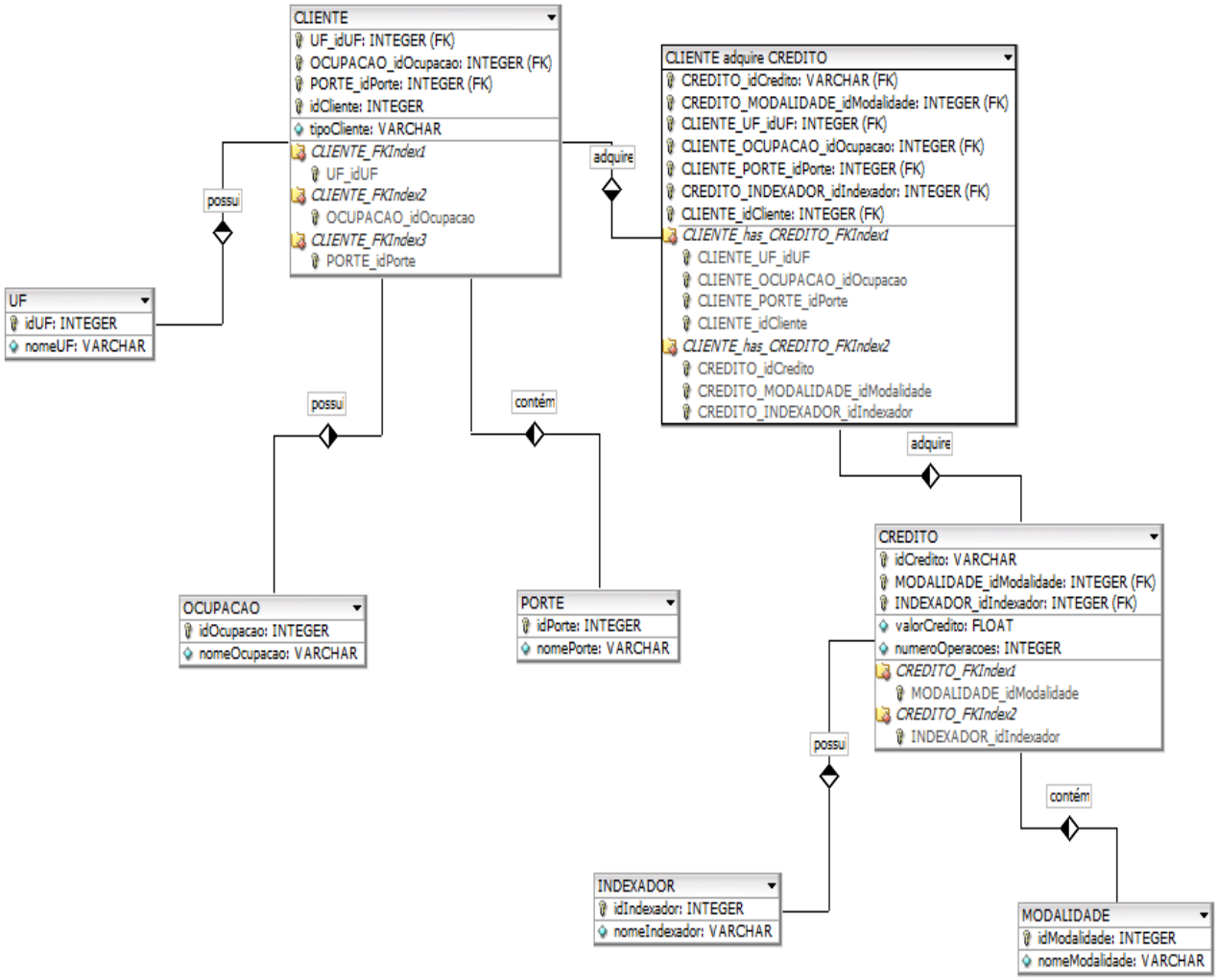
VALENTIM, Marta Lígia Pomim. Gestão da Informação e Gestão do Conhecimento em ambientes organizacionais. Tendências da Pesquisa Brasileira em Ciência da Informação, v. 1, n. 1, 2008.

SAYÃO, Luís Fernando. Modelos teóricos em ciência da informação-abstração e método científico. Ciência da informação, v. 30, p. 82-91, 2001.

WASSERMAN, Stanley; FAUST, Katherine. Social network analysis: methods and applications. Cambridge, England; New York: Cambridge University Press, 1994.

Apêndice A

Figura 4 - Modelagem conceitual sobre sete variáveis da base de dados



Fonte: autoria própria, com apoio do software DBDesigner.

DETECÇÃO DE VAZAMENTOS DE FLUIDOS DE FREIOS A AR EM VAGÕES DO TIPO GÔNDOLA ATRAVÉS DO SINAL ACÚSTICO: UM MODELO DE CLASSIFICAÇÃO DE FALHAS

AIR BRAKES FLUIDS LEAKAGE DETECTION IN GONDOLA WAGONS THROUGH ACOUSTIC SIGNALS: A FAULT CLASSIFIER MODEL

Jordana Lucia Reis[1], Flávio Miguel Varejão[1]
(1) Universidade Federal do Espírito Santo (UFES), Programa de Pós-Graduação em Informática, Av. Fernando Ferrari s/n, 29060-970, Vitória, ES, Brasil, jordana.reis@edu.ufes.br, flavio.varejao@ufes.br

Resumo
Este artigo apresenta um trabalho em andamento que utiliza um modelo de aprendizado de máquina para detectar sons de vazamento de fluidos de freios a ar em vagões de carga do tipo gôndola. Vagões gôndola transportam carga seca e possuem sistema de freio pneumático, dependendo diretamente de seus componentes de ar comprimido para uma frenagem adequada. Durante a inspeção de vagões, os vazamentos de ar comprimido são identificados a partir do som, reconhecido pelo ouvido humano. O fator de inovação deste trabalho é utilizar aprendizado de máquina para identificar vazamentos de ar comprimido a partir do som gravado. Os dados de entrada para o modelo são sinais acústicos capturados no formato *waveform*, durante acompanhamento da inspeção de vagões. No modelo proposto, o sinal acústico é processado no domínio tempo-frequência, para obtenção do espectrograma da melodia. Após o pré-processamento, os dados de espectrograma servem de entrada para o classificador *Random Forest*. Os resultados iniciais demonstraram que o modelo apresenta boas perspectivas para classificar os sons de vazamento de ar, mas é necessário aumentar o conjunto de dados para melhor avaliação do desempenho da proposta.

Palavras-chave: detecção de falhas; freios de vagões; sinal acústico; falhas de freios.

Abstract
This article presents a work in progress that applies a machine learning model to detect air brake fluid leak sounds in gondola-type cargo wagons. Gondola wagons carry dry cargo and have a pneumatic brake system, directly depending on their compressed air components for a properly braking. During wagon inspection, compressed air leaks are identified from the sound, recognized by the human ear. The innovation factor of this work is to apply machine learning to identify compressed air leaks from the recorded sound. The input data for the model are acoustic signals captured in waveform format, during monitoring of the inspection of wagons. In the proposed model, the acoustic signal is processed in the time-frequency domain to obtain the melody spectrogram. After pre-processing, the spectrogram data serves as input to the Random Forest Classifier. The initial results showed that the model has good prospects for classifying air leak sounds, but it is necessary to increase the dataset to better evaluate the proposal's performance.

Keywords: fault detection; wagons brakes, acoustic signal; braking fault.

1 Introdução

O uso do transporte ferroviário para escoamento de carga é muito utilizado na indústria, desde a Primeira Revolução Industrial. Este setor vem sofrendo evoluções e modernizações, mas continua suportando a indústria em seu objetivo de transportar grandes volumes de carga, de maneira mais barata e rápida, em comparação com outros meios de transporte disponíveis. Entretanto, ainda são tímidas as iniciativas de modernização e automatização dos processos de inspeção dos componentes ferroviários. Um exemplo disto é o procedimento de inspeção de freios dos vagões de carga. Tal procedimento é realizado por uma dupla de inspetores, que caminham ao longo de uma linha de vagões escutando atentamente em busca de identificar sons de vazamento de ar comprimido. Uma vez que o vazamento é identificado, é solicitada a retenção dos vagões após a descarga, para devida manutenção. Nem sempre é possível inspecionar todos os lotes de vagões, por exemplo, dependendo das condições climáticas, a inspeção é suspensa, para garantir a segurança do profissional que a executa, outra situação é caso o lote tenha chegado há mais de duas horas, todo o ar

comprimido já terá vazado e não há como identificar este tipo de falha. Existem outros fatores limitantes de inspeção de vazamento de fluidos de freios a ar, mas os exemplos citados são suficientes para demonstrar que o processo seguido atualmente ainda permite que muitos vagões circulem sem a devida verificação dos componentes de freios.

Em termos financeiros, o impacto do setor ferroviário, em geral, é significativo na estratégia logística da economia do Brasil, segundo dados da [ANTF 2021], apenas no ano de 2021, foram movimentadas 506,8 milhões de toneladas de cargas nas ferrovias brasileiras. Já em relação ao investimento em manutenção e reparos, a indústria ferroviária de forma global, investe um grande montante de recursos financeiros em manutenções e projetos de renovação de ferrovias e veículos ferroviários todos os anos, segundo estudo de [Xie et al. 2020].

Por outro lado, a média anual de acidentes ferroviários no Brasil, segundo apuração realizada por [Miguel 2020], estava em torno de 850 registros, até o ano de 2020. Neste contexto, obviamente, o sistema de frenagem de quaisquer veículos deve ser considerado uma parte crítica e importante para o bom funcionamento da máquina.

De acordo com [Jin et al. 2021], identificar falhas no sistema de frenagem dos trens é necessário para garantir a confiabilidade e saúde da composição ferroviária. Uma possível solução para detecção de falhas é aplicar a tecnologia para implementar o conceito de manutenção preditiva, isto é, monitorar um ativo para predizer quando este poderá apresentar falhas, substituindo-o antes do evento ocorrer. Além de garantir um sistema ferroviário mais confiável, a manutenção preditiva também reduz o custo de manutenções e renovações preventivas ou reativas, pois minimiza a indisponibilidade dos ativos através de planejamento e otimização da vida útil dos componentes.

O objetivo deste trabalho é propor um modelo de classificação de áudio capaz de detectar vazamento de ar comprimido. Os dados classificados devem permitir identificar se há falhas no sistema de freios a ar em vagões gôndola, de forma similar ao executado atualmente pelo profissional de inspeção de freios em campo.

O restante deste trabalho está organizado da seguinte forma: na seção 2, serão abordados os trabalhos relacionados ao assunto de aplicação de aprendizado de máquina para detecção de falhas em componentes de freios de vagões. Na seção 3, são descritos os procedimentos adotados para construção do modelo. Na seção 4, são apresentados os resultados alcançados neste trabalho até o momento. Finalmente, a seção 5 trata das conclusões parciais deste estudo.

2 Trabalhos Relacionados

Evitar os acidentes ferroviários, em específico os causados por falhas no sistema de freios, foi uma das motivações apresentadas para o desenvolvimento do trabalho de [Zhang et al. 2019], que implementou fusão e seleção de características multidimensionais, redução de dimensionalidade, e por fim, *Gradient Boosting Decision Tree* (GBDT) para diagnóstico das falhas. Segundo os autores, os resultados demonstraram excelente desempenho no reconhecimento de falhas em componentes de freios de vagões.

Já [Ye et al. 2020], desenvolveu uma pesquisa onde os sons produzidos pela frenagem dos vagões ferroviários serviram de entrada para um modelo de detecção de freios aplicados. Um dispositivo instalado ao lado da linha férrea, a princípio isolada para o estudo, gravava em formato *waveform* o som emitido pelas pás de freio em contato com as rodas dos trens no momento da aplicação e liberação dos freios. Assim, mapeou-se as amostras de som com freios aplicados e freios liberados. Aos dados sonoros, aplicou-se a Transformada Rápida de Fourier (FFT), seguida de representação de espectrograma e extração de características. Como o trabalho considerou 12 tipos diferentes de vagões, uma etapa de correspondência de padrão foi necessária, uma vez que os dados foram rotulados, aplicou-se o algoritmo Máquina de Vetores de Suporte (SVM) para classificação, alcançando 98% de acurácia. O estudo coletou dados em campo, compondo um conjunto de 66 amostras.

Em 2022, [Ji 2022] publicou um estudo focado no sistema de freios a ar para trens de alta-velocidade. A proposta do trabalho foi

isolar a falha identificada e diagnosticar o componente defeituoso, utilizando dados criados em uma plataforma para testes de freios de trens de alta-velocidade. A coleta de dados foi realizada através do uso de sensores para medir a pressão dos componentes do sistema de freios, especificamente válvula e cilindro. O conjunto de dados analisado era composto apenas de eventos de falha, contendo 70 amostras.

É importante ressaltar que a revisão de literatura realizada é não-exaustiva, mas até o momento, não foram encontradas pesquisas que utilizem dados de sinais acústicos para detectar falhas de vazamentos de fluidos de freios de vagões gôndola, objeto de estudo neste trabalho.

3 Procedimentos Metodológicos

A metodologia aplicada neste trabalho obedece a seguinte ordem: coleta de dados sonoros em campo; pré-processamento dos dados; desenvolvimento e treinamento de classificador; e teste e avaliação de desempenho do classificador.

3.1 Coleta de Dados

O procedimento de coleta do conjunto de dados foi realizado em campo, durante acompanhamento de rotina de inspeção de freios de vagões, na companhia Vale S.A. A captura do som foi feita através da utilização de *smartphone,* utilizando um aplicativo gratuito para gravação de áudio em waveform, chamado waveEditor, disponível nas lojas dos sistemas operacionais destes dispositivos. A exemplo da inspeção realizada por humanos, a distância do dispositivo de gravação ao vagão, foi estabelecida em aproximadamente 1 metro, salvo algumas exceções devido a irregularidades do terreno.

Os componentes do sistema de freios avaliados são: cilindros, válvulas e mangotes. Estes são os componentes nos quais os vazamentos de ar comprimido podem ser identificados.

Foram inspecionados três lotes de vagões, contendo em média, 84 vagões cada lote. Até o momento da escrita deste artigo, foram realizadas coletas sobre 13 eventos de vazamento de ar comprimido, fluido do sistema de freios dos vagões gôndola. De maneira análoga, foram gravadas 13 situações de condição normal onde não há som de vazamento de ar. Cada gravação dura aproximadamente 5 segundos.

Após a captura do som, cada amostra foi rotulada em condição de falha e condição normal, especificada a linha de inspeção e o lado do vagão onde foi realizada a gravação. Cada amostra foi nomeada de acordo com estas informações. Por exemplo, o arquivo de áudio fe13_2022_07_15_163536544.wav, refere-se à captura de áudio de uma situação de falha (f), identificada ao percorrer o lado esquerdo (e) do vagão, cujo lote foi inspecionado na linha de inspeção de número 13, seguido de data e hora da coleta.

A Figura 2 é uma imagem capturada durante a inspeção de freios de vagões, onde é possível observar dois vagões gôndola, carregados de minério de ferro no pátio de inspeção, pelo lado direito dos vagões, destacados pelo retângulo de cor branca, os componentes analisados nesta pesquisa, onde é possível identificar vazamentos de ar comprimido.

Figura 1 - Visualização de componentes de freios inspecionados na pesquisa. (Dados da Pesquisa, 2022)

3.2 Pré-processamento

Um sinal acústico pode ser representado como uma forma de onda em função do tempo, representando a intensidade do som em cada instante do tempo.

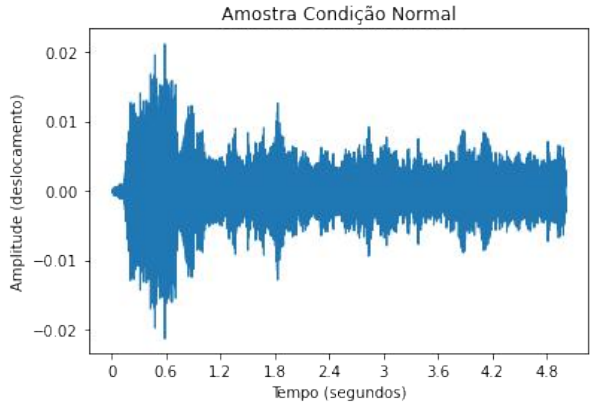


Figura 2 - Representação gráfica de amostra de som de condição normal em amplitude x tempo (Dados da Pesquisa, 2022)

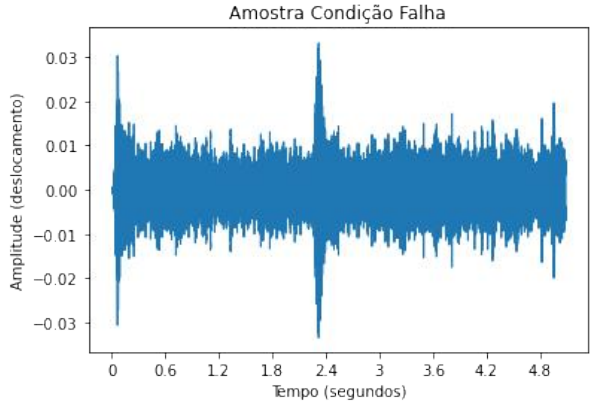


Figura 3 - Representação gráfica de amostra de som de condição de falha em amplitude x tempo (Dados da Pesquisa, 2022)

Para cada gravação, aplicou-se o seccionamento em arquivos de 1 segundo, resultando em 153 arquivos, e posterior aquisição dos dados no domínio tempo-frequência, utilizando a Transformada de Fourier de Tempo Curto (STFT), gerando matrizes para treinamento e teste do modelo de classificação. A biblioteca librosa foi utilizada como suporte no processo de aplicação do método STFT, sendo que os parâmetros utilizados foram estabelecidos como: 1) *n_fft* igual a 2048, para o tamanho da FFT. 2) *hop_length*, igual a 1024, para o tamanho do salto. 3) *window*, foi a lista resultante da execução função Blackman-Harris que recebeu o parâmetro número de pontos de saída igual a 1024. Os demais parâmetros foram mantidos exatamente como o padrão da biblioteca.

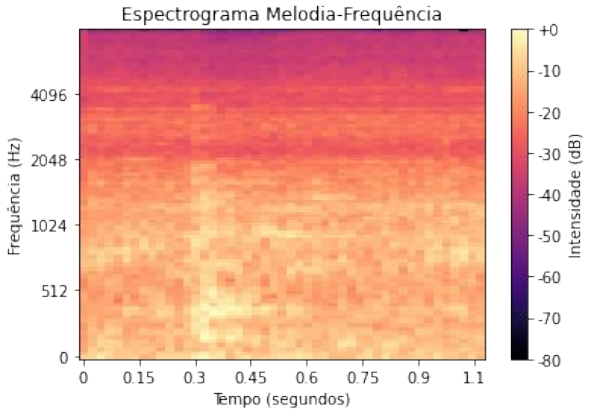


Figura 4 - Exemplo de espectrograma de melodia de uma amostra de falha (Dados da Pesquisa, 2022)

Como seleção de características, aplicou-se o espectrograma de melodia com taxa de amostragem igual a 44.100, dividido em 100 bandas, compostas de 47 valores de espectrograma de melodia. Em seguida, os dados dos arquivos foram normalizados em valores entre 0 (zero) e 1 (um).

Como os áudios originais foram gravados com aproximadamente 5 segundos, alguns arquivos continham alguns milissegundos a mais e alguns continham alguns milissegundos a menos. Desta forma, quando aplicado o seccionamento em 1 segundo de duração, algumas seções ficaram menores do que 1 segundo. Para garantir a padronização, os arquivos menores que 1 segundo, no total 26, foram desconsiderados. Portanto, o conjunto contou com 127 arquivos, de 47 colunas e 100 linhas, linhas tais geradas a partir das 100 bandas do espectrograma. Destes arquivos, 65 são referentes a amostras de áudio em condições normais e 62 referem-se a amostras de áudio capturado em condição de falha. Assim, foi formada uma base para alimentar o modelo com 12.700 amostras, rotuladas em duas classes, condição normal e condição de falha, de acordo com a nomenclatura do arquivo original.

3.3 Modelo

O modelo escolhido para os testes foi o *Random forest classifier* (RFC), pois apresentou melhor resultado em validação cruzada repetida 10 vezes, quando comparado com os modelos K-vizinhos mais próximos, Árvore de decisão e Máquina de vetores de suporte.

Os dados foram separados em conjuntos de treino e teste, sendo que 20% do conjunto foi utilizado para teste, isto é, 2.540 entradas, e o restante para treino. O conjunto foi separado em treino e teste de forma randômica, utilizando a semente igual a sete (7). O número de estimadores utilizado como parâmetro para a execução do modelo foi igual a um mil (1000). Os resultados apresentados na próxima seção são parciais e foram obtidos com um conjunto reduzido de amostras, apresentando informações considerando o conjunto de testes, que não foi utilizado no processo de treinamento do modelo.

4 Resultados

A acurácia alcançada pelo modelo foi de 0,73% e F1-score acima de 70% para ambas classes. Entretanto, algumas condições de falha foram incorretamente rotuladas como condição normal pelo modelo, o que pode ser prejudicial para o propósito de auxiliar o processo de inspeção dos freios de vagão.

É importante ressaltar que este trabalho está em andamento e uma evolução já identificada é avaliar outras formas seleção de características para fornecer como entrada ao modelo, as características que ofereçam informações de qualidade para melhorar o desempenho do modelo proposto, assim como utilizar a Transformada Rápida de Fourier (FFT) diretamente. Além disto, o conjunto de dados está em processo de formação, sendo incrementado a cada acompanhamento de inspeção em campo, processo pelo qual são capturadas as amostras de áudio.

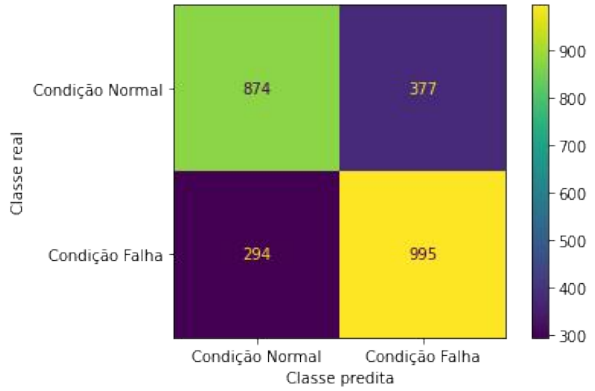


Figura 5 - Matriz de confusão (Dados da Pesquisa, 2022)

5 Considerações Finais

Este trabalho apresentou uma abordagem para utilizar dados de sinal acústico para identificar vazamentos de fluidos de freios de vagão, e consequente detecção de falhas no sistema de frenagem de vagões de transporte de carga. Os resultados iniciais indicam que é possível utilizar o som gravado em aplicativo gratuito de telefone móvel para alimentar um modelo de classificação inteligente, capaz de auxiliar o processo de inspeção de freios de vagões gôndola.

Na evolução do desenvolvimento deste trabalho, pretende-se avaliar outros tipos de espectrogramas para representar o sinal acústico, comparando ao desempenho alcançado neste estudo. Além disso, há espaço para trabalhar mais detalhes do áudio de entrada, eventualmente capturando o sinal com outros tipos de equipamentos, que embora possam restringir a utilização da proposta, possam trazer mais informações sobre o dado pela qualidade da gravação. Assim também, evoluir na exploração e mapeamento de frequências, amplitudes, e demais atributos que contribuam para direcionar a criação de um modelo mais assertivo.

Referências

ANTF. *Informações gerais: o setor ferroviário de carga brasileiro*. Associação Nacional dos Transportadores Ferroviários, 2021. Disponível em: https://www.antf.org.br/informacoes-gerais/. Acesso em: 13 nov. 2022.

JI, Hongquan. *Optimization-based incipient fault isolation for the high-speed train air brake system*. IEEE Transactions on Instrumentation and Measurement, 71, 2022. Disponível em: https://ieeexplore.ieee.org/document/9666849. Acesso em: 13 nov. 2022.

JIN, Yongze *et al*. *Fault diagnosis of brake train based on multi-sensor data fusion*. Sensors, 21, 2021. Disponível em: https://www.mdpi.com/1424-8220/21/13/4370. Acesso em: 13 nov.2022.

MIGUEL, Daniel. *Acidentes ferroviários no brasil: análise comparativa com a união europeia*. *In*: Congresso ANPET, 34, 2020. Disponível em: https://www.anpet.org.br/anais34/documentos/2020/Aspectos Econômicos Sociais Políticos e Ambientais do Transporte/Gestão do Transporte Ferroviário e Hidroviário/4_220_AC.pdf. Acesso em 13 nov. 2022.

XIE, Jiawei *et al*. *Systematic literature review on data-driven models for predictive maintenance of railway track: implications in geotechnical engineering*. Geosciences (Switzerland), 10:1–24, 2020. Disponível em: https://www.mdpi.com/2076-3263/10/11/425. Acesso em: 13 nov. 2022.

ZHANG, Meng; LIU, Zhen; Dang, Xinyue. *Fault diagnosis on train brake system based on multi-dimensional feature fusion and gbdt enhanced classification*. 2018 International Conference on Intelligent Rail Transportation (ICIRT), Singapura, 2019. Disponível em: https://ieeexplore.ieee.org/document/8641607. Acesso em: 13 nov. 2022.

YE, Yuling; ZHANG, Jun; LIANG, Hengda. *An acoustic-based recognition algorithm for the unreleased braking of railway wagons in marshalling yards*. IEEE Access, 8:120295–120308, 2020. Disponível em: https://ieeexplore.ieee.org/document/9129748. Acesso em: 13 nov. 2022.

DIAGNÓSTICO DE FALHAS: UMA REVISÃO E ANÁLISE DE DADOS DE VIBRAÇÃO E SUAS APLICAÇÕES

FAULT DIAGNOSIS: A REVIEW AND ANALYSIS OF VIBRATION DATA AND ITS APPLICATIONS

Igor Varejão[1], Alexandre Rodrigues Loureiros[2], Thiago Oliveira-Santos[3], Flávio Varejão[4]
(1) UFES, Espirito Santo, igor.varejao@edu.ufes.br
(2) UFES, Espirito Santo, arodrigues.ufes@gmail.com
(3) UFES, Espirito Santo, todsantos@inf.ufes.br
(4) UFES, Espirito Santo, flavio.varejao@ufes.br

Resumo
Empresas do ramo industrial geralmente têm grandes investimentos em equipamentos modernos de produção, bem como altos custos de manutenção dessas unidades. A identificação rápida e precisa de falhas e problemas nos equipamentos industriais contribui de forma crucial para reduzir os custos de manutenção e melhorar a confiabilidade na produção. O diagnóstico de falhas consiste no monitoramento do funcionamento de um equipamento de modo a identificar a ocorrência de uma falha. Com o aumento do número de sensores instalados a bordo dos equipamentos, os dados adquiridos por eles têm sido mais usados para monitorar os estados desses equipamentos e diagnosticar suas falhas ou mal funcionamento. Avanços nas pesquisas na área de Inteligência Artificial, em especial, na área de Aprendizado de Máquina, fornecem meios para aumentar a confiabilidade de sistemas inteligentes de diagnóstico de falhas que resultam em um desempenho mais confiável dos equipamentos e, por consequência, da indústria. Este artigo apresenta uma análise geral, assim como, uma revisão, sobre os dados de vibração que têm sido usados em diversos trabalhos de pesquisa nos últimos 30 anos, aponta um problema comum no uso desses dados, apresenta o que precisa ser feito para resolvê-lo e como a comunidade acadêmica pode contribuir para esta solução.

Palavras-chave: Diagnóstico de Falhas; Aprendizado de Máquina; Análise de Dados de Vibração; Bias de Similaridade.

Abstract
Companies in the industrial sector generally have large investments in modern production equipment, as well as high maintenance costs for these units. Fast and accurate detection of failures and problems in industrial equipment makes a crucial contribution to reducing maintenance costs and improving confidence in production. Fault diagnosis consists of monitoring the operation of equipment in order to identify the occurrence of a failure. With the increase in the number of sensors installed on board in equipment, they have been more used to monitor the status of these equipment and diagnose their failures or malfunctions. Advances in research in the area of Artificial Intelligence, especially in the area of Machine Learning, provide ways to increase the reliability of intelligent fault diagnosis systems and result in a more reliable performance of equipment and industry. This article presents an overview of the vibration data that have been used in several works in the last 30 years, points out a common problem in the use of these data, presents what needs to be done to solve it and how the academic community can contribute to this solution.

Keywords: Fault Diagnosis; Machine Learning; Vibrational Data Analysis; Similarity Bias.

1. Introdução

O diagnóstico inteligente de falhas em equipamentos industriais consiste em aplicar técnicas de inteligência artificial para identificar a ocorrência de falhas e qual o tipo de falha que ocorreu. Este é um problema categorizado como de classificação, no qual o modelo é induzido para escolher uma ou mais classes dentre um conjunto de possíveis classes. De modo geral, uma das classes do problema é a classe normal que apresenta exemplos do equipamento funcionando sem a presença de qualquer falha. As demais classes apresentam exemplos de diferentes tipos de falhas que

podem ocorrer durante a operação do equipamento.

Técnicas de Inteligência Artificial aplicadas no diagnóstico inteligente de falhas tipicamente utilizam dados coletados por sensores ao longo do tempo de operação dos equipamentos industriais. Os sensores mais comumente utilizados são os acelerômetros, que capturam a magnitude da vibração em um determinado instante. A coleta de dados pode ser contínua ou ocorrer em intervalos regulares de tempo. Cada medição pode conter dados de um único instante no tempo ou de um intervalo de tempo. De modo geral, a sequência temporal na qual os dados são coletados têm relevância para o diagnóstico das falhas. As curvas temporais representadas pela sequência de dados indicam padrões de normalidade e falha.

Existe uma literatura abundante de trabalhos que utilizam dados de vibração para diagnóstico inteligente de falhas em equipamentos industriais os quais utilizam dados de vibração coletados em ensaios laboratoriais ou de equipamentos em operação em campo. Técnicas de Inteligência Artificial são usadas para que o sistema aprenda modelos capazes de realizar o diagnóstico de falhas a partir da coleta de amostras dos dados de vibração do equipamento em operação. Os resultados reportados nestes trabalhos são de modo geral bons ou excelentes, trazendo uma grande esperança de que essas técnicas possam ser aplicadas no dia a dia das indústrias.

Este artigo, contudo, indica que o uso dessas técnicas no dia a dia ainda está distante. Ele aponta e analisa aprofundadamente um problema comum nos trabalhos reportados na literatura, relacionado à forma como os dados são utilizados e na metodologia de avaliação utilizada. O artigo também propõe uma maneira de eliminar esse problema, embora isso envolva um grande esforço da comunidade científica e industrial.

Além desta introdução, este artigo é organizado da seguinte forma: a Seção 2 apresenta uma visão geral e breve das principais técnicas de Inteligência Artificial usadas com dados de vibração em equipamentos industriais para o diagnóstico inteligente de falhas. A Seção 3 apresenta as bases de dados mais frequentemente utilizadas nestes trabalhos, resumindo suas caraterísticas mais relevantes. A Seção 4 identifica o problema comum existente nestes trabalhos e os categorizam em diferentes subtipos. A Seção 5 apresenta a nossa proposta para a minimização desses problemas. Finalmente, a seção 6 apresenta as nossas conclusões.

2. Diagnóstico de Falhas em Equipamentos Industriais

A maior parte das técnicas de aprendizado usadas para diagnóstico de falhas em equipamentos industriais é de aprendizado supervisionado, requerendo assim que os exemplos da base indiquem se representam uma condição de operação normal do equipamento ou uma condição que representa uma falha específica. O tipo de dado da base de exemplos mais utilizado no diagnóstico de falhas é o sinal de vibração do equipamento.

2.1. Domínio dos Dados

O sinal de vibração é originalmente extraído como uma série-temporal no domínio do tempo, porém, para a realização do diagnóstico na prática se aplicam transformações no sinal para obter seu espectro no domínio da frequência, o qual permite a identificação de padrões não evidentes no domínio do tempo.

2.1.1. Tempo

O sinal temporal é obtido a partir da coleta de medições de sensores do tipo acelerômetro localizados em posições específicas do equipamento ao longo de um período de tempo. Cada medição consiste da amplitude da vibração em um determinado instante. O sinal extraído no domínio do tempo pode ser utilizado diretamente como entrada nos classificadores, neste caso, cada medição corresponde a uma entrada, ou como é mais comum, são extraídas características do sinal as quais servem como entrada do classificador. Exemplos de características comuns são estatísticas como média da amplitude, desvio padrão, curtose, pico máximo e rms (*root mean square*) do sinal. A figura 1 apresenta um exemplo de

um sinal de vibração coletado no domínio do tempo.

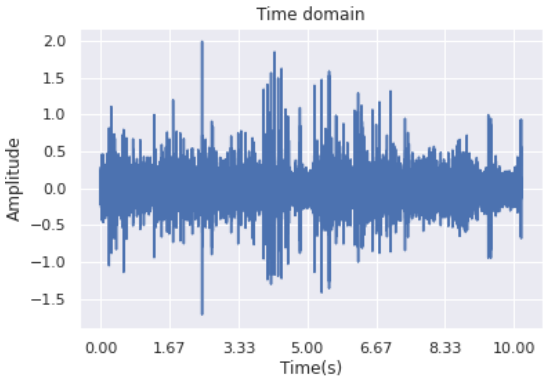


Figura 1 - Sinal de vibração no domínio do tempo

2.1.2. Frequência

Em situações do mundo real, máquinas complexas com muitos componentes geram uma gama variada de vibrações. Dessa forma, na prática da engenharia é difícil utilizar os sinais no domínio do tempo para analisar a situação do equipamento. Engenheiros especialistas lançam mão da transformação do sinal para o domínio da frequência, onde pode se identificar picos de amplitude em frequências típicas de defeitos. A transformação pode ser realizada facilmente pela Transformada Rápida de Fourier (TRF). A figura 2 mostra um exemplo de sinal transformado para o domínio da frequência de um defeito de desbalanceamento de bombas centrífugas. Note o pico existente na frequência de rotação do equipamento.

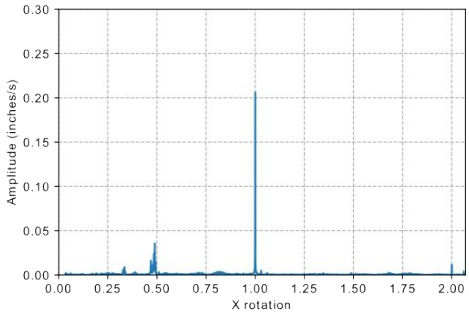


Figura 2 - Sinal de vibração no domínio da frequência

2.2. Técnicas de Aprendizado de Máquina

Várias técnicas de aprendizado de máquina têm sido usadas neste problema, envolvendo desde a extração e seleção de características a variados algoritmos de classificação.

2.2.1. Tradicionais

Métodos de classificação tradicionais como k vizinhos mais próximos (WANG, 2016), máquinas de vetores de suporte (RAUBER et al., 2021) e florestas de árvores de decisão (MELLO et al., 2022) têm sido aplicados com sucesso no diagnóstico inteligente das falhas. Normalmente esses métodos são aplicados a partir de características estatísticas definidas com o auxílio de conhecimento especializado sobre o funcionamento do equipamento específico em foco no problema. A definição dessas características é a etapa mais relevante e trabalhosa para a aplicação dos métodos. Uma revisão ampla sobre a aplicação desses métodos nesse problema pode ser encontrada em (LIU et al., 2018).

2.2.2. Aprendizado Profundo

Mais recentemente, técnicas de aprendizado profundo (MELLO et al., 2022) têm obtido resultados ainda mais promissores nesta tarefa. Uma grande vantagem destas técnicas é extrair diretamente as características relevantes a partir da entrada do sinal bruto no domínio do tempo ou da frequência, dispensando a necessidade de uso de conhecimento especializado prévio, o que proporciona maior capacidade de generalização da aplicação.

As técnicas de aprendizado profundo têm sido usadas tanto como extratores automático de características, neste caso, as características extraídas são depois usadas como entradas para classificadores tradicionais, como também como métodos de classificação, neste caso uma camada adicional de classificação é adicionada para que a própria rede neural profunda possa realizar a classificação.

Uma ampla revisão sobre a aplicação desses métodos de aprendizado profundo para diagnóstico inteligente de falhas baseada em sinais de vibração pode ser encontrada em (ZHANG et al., 2020).

3. Análise de Dados de Vibração

Os métodos de aprendizado de máquina dependem fundamentalmente dos dados usados como entrada. Dessa forma, torna-se necessário garantir que os métodos

usados recebam dados de qualidade e numerosos, garantindo uma diversidade suficiente entre os dados para que o modelo possa abstrair as informações e aprender a generalizar a tarefa de diagnóstico. Contudo, apesar da evolução na aquisição dos dados em equipamentos o processo ainda é demorado e caro dado que a máquina leva um longo tempo para falhar. Sendo assim, instituições ao redor do mundo realizaram experimentos em ambientes controlados induzindo falhas em máquinas ou acelerando o processo para, assim, proporcionar à comunidade *datasets* para treinarem seus modelos. Todavia, cada experimento tem sua particularidade que deve ser levada em conta, como o tipo de equipamento, a frequência de rotação, e as condições de operação. Dado isso, apresentamos a seguir os principais *datasets* públicos no contexto de equipamentos industriais dividido por graus de similaridade.

3.1. Mesmo Equipamento, Mesmas Condições e Diferentes Estados

Há *datasets* em que os sinais de diferentes classes são adquiridos em contextos iguais, ou seja, o sinal de normal e o de falhas é adquirido a partir de um único equipamento usando-se as mesmas condições, por exemplo, a mesma frequência de rotação e a mesma carga a qual o equipamento é submetido. O *dataset* disponível desse tipo é o *Intelligent Maintenance Systems* (IMS)(QIU et al., 2003).

Os dados do IMS foram coletados a partir de 3 experimentos *run-to-failure*, cada um utilizando um equipamento de teste composto por 4 rolamentos atrelados a uma haste, a qual estava acoplada a um motor. Assim, o instituto acelerou o processo de degradação dos rolamentos e, dessa forma, conseguiu obter 4 estados: normal, falha da pista interna, falha da pista externa e falha da roldana.

3.2. Mesmo Equipamento, Diferentes Condições

O CWRU (*Case Western Reserve University*) (CWRU, 2021) é provavelmente o *dataset* mais referenciado na literatura no contexto de detecção de falhas em equipamentos. Os dados disponibilizados neste *dataset* foram adquiridos por meio de um equipamento de teste que consiste em um motor elétrico no qual falhas foram simuladas e testadas, incluindo falha na pista interna, falha na pista externa e falha na bola do rolamento, assim como o estado normal da máquina.

Outro *dataset* nesta categoria é o MFPT (*Mechanical Failures Prevention Group*)(MFPT, 2012) *bearing dataset*. O MFPT é composto por 4 estados de rolamento, todos obtidos com a mesma frequência de rotação, utilizando o mesmo equipamento de teste. Os estados englobados são as condições dos rolamentos normais, falha na pista interna e falha na pista externa.

Além desses, tem-se o *dataset* MAFAULDA (*Machinery Fault Database*) (MAFAULDA, 2021) disponibilizado pelo SMT(*Signals, Multimedia, and Telecommunications Laboratory*) da UFRJ composto por experimento feitos em um simulador de falhas de máquinas basicamente composto de rolamentos conectados por uma haste, atrelada a um motor. Nesse experimento foram simulados seis estados diferentes: estado normal, falha de desbalanceamento, desalinhamento vertical e horizontal e falha no rolamento externo e interno.

3.3. Mesmo Tipo de Equipamento, Diferentes Equipamentos, Diferentes Condições

O *dataset* ESPset (*Eletrical Submersal Pump datase*t) executa a coleta dos dados de vibração em experimentos realizando em bombas centrífugas submersas (BCS), equipamento de grande importância na indústria do petróleo. O *dataset* contém dados de vibrações de 12 BCSs diferentes, nas quais uma BCS é testada com diferentes frequências de rotação e monitorada por 36 acelerômetros localizados em diferentes posições da BCS. Foram identificados 5 estados no *dataset* inteiro, o sinal normal, falha no sensor/acelerômetro, desalinhamento, roçamento e desbalanceamento, totalizando 6032 sinais de vibração.

4. Bias de Similaridade

Uma parte essencial do aprendizado de máquina é saber avaliar o desempenho do algoritmo de tal forma que o resultado apresentado seja consistente e desempenhe de acordo quando exposto a dados nunca vistos do mundo real. Entretanto, muitos trabalhos de diagnósticos inteligentes de falhas baseadas na análise de sinais de vibração pecam em garantir essa consistência na amostragem e seus trabalhos ficam sensíveis ao chamado **bias de similaridade** (RAUBER et al., 2021). Isso ocorre porque a aquisição de dados é feita por meio de séries temporais as quais são divididas em janelas para gerar instâncias diferentes para o treinamento dos classificadores. Dessa forma, deveria ser garantido que a divisão entre treino e teste não contém padrões extraídos de um mesmo sinal em períodos de tempo similares já que, caso isso ocorra, a tarefa de classificação se torna relativamente simples. É importante ressaltar que, mesmo que o pesquisador realize a amostragem meticulosamente tentando evitar o bias de similaridade, há casos que o *dataset* não permite que o bias seja totalmente evitado. Tendo em vista esse cenário há vários graus de bias que podem ocorrer.

4.1. Máximo

Quando os conjuntos de treino e teste levam em conta sinais de um mesmo equipamento testado com as mesmas condições, a pesquisa fica extremamente enviesada e o bias de similaridade está em seu grau máximo uma vez que, como apontado acima, a tarefa de classificação se torna trivial dado que há uma grande dependência entre conjunto de treino e teste. Conforme ilustrado em (RAUBER et al., 2021), esse caso acaba ocorrendo na maioria dos artigos de aprendizado de máquina aplicado ao diagnóstico de falhas baseada em dados de vibração, os quais rotineiramente reportam taxa extremamente elevadas de acerto, beirando à 100% de acurácia. (RAUBER et al., 2021) realiza um experimento usando validação cruzada aninhada na base de dados CWRU no qual exemplos dos conjuntos de treino e teste provém das mesmas séries temporais e, como esperado, atingem resultados muito próximos a 100% de acurácia. Por conseguinte, os métodos usados nesses artigos tendem a não funcionar até mesmo em equipamentos utilizados na pesquisa mas em condições diferentes.

4.2. Intermediário

Pesquisas que utilizam técnicas de amostragem nos quais o conjunto de treino possui instâncias de sinais temporais obtidos em determinadas condições e o conjunto de teste possui instâncias de sinais temporais oriundos de outras condições reduzem o bias de similaridade, porém quando os dados são obtidos a partir de um mesmo equipamento e as condições não variam de forma significativa, acabam por manter um grau intermediário de bias de similaridade. (RAUBER et al., 2021) realizou dois outros experimentos com a base de dados CWRU no qual os conjuntos de treino e teste foram separados de forma a garantir que os exemplos de teste não provenham das mesmas séries temporais utilizadas para obter os exemplos de treino.

No primeiro deles, a diferença entre os exemplos dos conjuntos de treino e teste foi alcançada usando diferentes cargas no equipamento. Neste experimento as diferenças de carga não tinham muito impacto na forma das séries temporais e os resultados ficaram próximos de 95% de acurácia, indicando um bias de similaridade intermediário, porém ainda alto.

No segundo experimento, as diferenças foram obtidas com variação de carga e variação da severidade da falha. Os resultados caíram significativamente, para próximo de 50% de acurácia, mostrando o decaimento significativo do desempenho dos métodos avaliados quando se reduz o bias .

4.3. Mínimo

Quando o conjunto de treino e teste é amostrado de tal forma que os dados do conjunto de teste são adquiridos de um equipamento distinto dos equipamentos usados para obter os exemplos de treino, preferencialmente sob várias condições de uso, o bias de similaridade é mínimo e, neste caso, os resultados obtidos são mais prováveis de serem generalizáveis.

5. Integração e Ampliação das Bases de Dados

Mesmo se tomando cuidados para minimizar o bias de similaridade, a capacidade de generalização dos resultados continua questionável, visto que, na prática, as condições reais são bem variadas e normalmente distintas das condições usadas nos experimentos.

Uma maneira de avançar na solução desse problema seria montar uma grande base de dados composta pela integração das bases de dados públicas em uma única base usada para treinamento e teste dos classificadores. Classificadores com bons resultados teriam maior possibilidade de generalização considerando que o conjunto de dados usado para treinamento é bem mais variado e distinto do conjunto usado no teste. Atualmente, temos investido na formação dessa base de dados e avaliado o impacto de usá-la junto aos métodos de diagnóstico inteligente de falhas.

Um projeto mais desafiador e promissor nesta direção seria criar uma base de dados integrada e ampliada coletando sinais em diferentes condições de diferentes tipos de equipamentos para várias tarefas de diagnóstico de falhas, como o *dataset* ImageNet (DENG et al., 2009) da comunidade de visão computacional. Dessa forma, haveria dados suficientes para que a divisão entre treino e teste fosse feita sem qualquer *bias* de similaridade. Este é um projeto desafiador que vai requisitar um grande esforço da comunidade científica e industrial.

6. Considerações Finais

Até o momento da publicação desse artigo não há *dataset* público que utilize diferentes tipos de equipamentos com diferentes equipamentos para aquisição de dados de vibração. Dessa forma, torna-se um esforço necessário a divulgação para a comunidade acadêmica um *dataset* integrado de dados de vibração, como proposto nesse artigo, para que futuros trabalhos sejam feitos de forma que a amostragem seja montada de modo que vários tipos de equipamentos sejam utilizados e que, no conjunto de teste, haja tipos de equipamentos não encontrados no conjunto de treino. Isto posto, não haveria *bias* de similaridade na pesquisa e, portanto, caso algum artigo reporte bons resultados satisfazendo essas condições, indicaria que o uso dessas técnicas no dia a dia é bastante viável.

Referências

CWRU. **Case Western Reserve University Bearing Dataset Center**. Disponível em: <https://engineering.case.edu/bearingdatacenter>. Acesso em: 17 set. 2022.

DENG, J. et al. **ImageNet: A large-scale hierarchical image database**. 2009 IEEE Conference on Computer Vision and Pattern Recognition. **Anais**... Em: 2009 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. jun. 2009.

LIU, R. et al. Artificial intelligence for fault diagnosis of rotating machinery: A review. **Mechanical Systems and Signal Processing**, v. 108, p. 33–47, 1 ago. 2018.

MAFAULDA. **Machinery Fault Database**. Disponível em:<http://www02.smt.ufrj.br/~offshore/mfs/page_01.html#SEC1>. Acesso em: 17 set. 2022.

MELLO, L. H. S. et al. Ensemble of metric learners for improving electrical submersible pump fault diagnosis. **Journal of Petroleum Science and Engineering**, v. 218, p. 110875, 1 nov. 2022.

MFPT. **Fault Data Sets - Society For Machinery Failure Prevention Technology**. Disponível em: <https://www.mfpt.org/fault-data-sets/>. Acesso em: 17 set. 2022.

QIU, H. et al. Robust performance degradation assessment methods for enhanced rolling element bearing prognostics. **Advanced Engineering Informatics**, Intelligent Maintenance Systems. v. 17, n. 3, p. 127–140, 1 jul. 2003.

RAUBER, T. W. et al. An experimental methodology to evaluate machine learning methods for fault diagnosis based on vibration signals. **Expert Systems with Applications**, v. 167, p. 114022, 1 abr. 2021.

WANG, D. K-nearest neighbors based methods for identification of different gear crack levels under different motor speeds and loads: Revisited. **Mechanical Systems and Signal Processing**, v. 70–71, p. 201–208, 1 mar. 2016.

ZHANG, S. et al. Deep Learning Algorithms for Bearing Fault Diagnostics—A Comprehensive Review. **IEEE Access**, v. 8, p. 29857–29881, 2020.

DIREITOS AUTORAIS RELACIONADOS À MEMÓRIA INSTITUCIONAL E ARTÍSTICA DO TRIBUNAL DE JUSTIÇA DO DISTRITO FEDERAL E DOS TERRITÓRIOS

COPYRIGHT RELATED TO THE TJDFT'S INSTITUTIONAL AND ARTISTIC MEMORY

Rosilene Paiva Marinho de Sousa[1] Maison Roberto M. Gonçalves[2], Diego José Macedo[3], Milton Shintaku[4]

(1) Instituto Brasileiro de Informação em Ciência e Tecnologia, SAUS Quadra 5, Lote 6, Bloco H - (61), rosilenesousa@ibict.br

(2) Instituto Brasileiro de Informação em Ciência e Tecnologia, SAUS Quadra 5, Lote 6, Bloco H - (61), maisongoncalves@ibict.br

(3) Instituto Brasileiro de Informação em Ciência e Tecnologia, SAUS Quadra 5, Lote 6, Bloco H - (61), diegomacedo@ibict.br

(4) Instituto Brasileiro de Informação em Ciência e Tecnologia, SAUS Quadra 5, Lote 6, Bloco H - (61), shintaku@ibict.br

Resumo

As tecnologias de informação e comunicação têm contribuído para a preservação da memória institucional de diversos órgãos do poder público. Diante disso, desperta-se a preocupação com os direitos autorais e de imagem dos documentos digitais que compõem o acervo, considerando sua importância na garantia de direitos morais e patrimoniais do autor em obras literárias artísticas e científicas. Nesse sentido, este trabalho tem por objetivo apresentar uma análise dos direitos autorais dos documentos constantes no acervo digital do Memorial do TJDFT, apresentados no sistema de informação implementado com o software livre Omeka. Para isso, aborda o sistema de informação baseado no Omeka, apresentando suas principais características e importância para preservação da memória institucional. Apresentam-se os principais aspectos que envolvem a proteção de direitos autorais e sua distinção com a proteção ao direito de imagem. Como metodologia, adotar-se-á pesquisa qualitativa, exploratória e bibliográfica. Como resultado, examinam-se os direitos autorais da tipologia documental que compõem o acervo digital do TJDFT apresentado no próprio sistema. Conclui-se que o sistema de informação baseado no Omeka atende aos requisitos legais de proteção do acervo digital, pois permite a inclusão dos termos, embora haja uma necessidade de maiores subsídios sobre a regulamentação da transferência de direitos autorais e de imagem de fotos de prédios e pessoas que compõem o acervo do TJDFT.

Palavras-chave: Direito autoral; Direito de imagem; Fotografia; Obra de arte; TJDFT.

Abstract

Information and communication technologies have contributed to the preservation of the institutional memory of various organs of public power. Therefore, concern has been aroused with the copyright and image of the digital documents that make up the collection, considering their importance in guaranteeing the author's moral and property rights in artistic and scientific literary works. In this sense, this work aims to present an analysis of the copyright of the documents contained in the digital collection of the TJDFT Memorial, presented in the information system implemented with the free software Omeka. For this, the information system based on Omeka is approached, presenting its main characteristics and importance for the preservation of institutional memory. It presents the main aspects that involve the protection of copyright and its distinction with the protection of image rights. Qualitative, exploratory and bibliographic research will be adopted as a methodology. As a result, we examine the copyright of the documentary typology that make up the digital collection of the TJDFT presented in the system itself. It is concluded that the information system based on Omeka meets the legal requirements for the protection of the digital collection, by allowing the inclusion of terms, although there is a need for greater subsidies on the regulation of copyright and image transfer of photos of buildings and people who make up the TJDFT collection.

Keywords: Metadata; Copyright; Image rights; Photography; Work of art; TJDFT.

1. Introdução

A história do Tribunal de Justiça do Distrito Federal e dos Territórios (TJDFT), por ter sede na capital, data do Brasil Império, em que a capital era Salvador. Entretanto,

como destacado no portal do tribunal[1], sua história atual data de 1960, com a transferência da capital do Brasil para Brasília por meio da Lei 3.754, de 14 de abril de 1960, que organiza o poder judiciário federal na capital.

Com uma história tão rica, o TJDFT, nascido em dois andares do Ministério da Justiça, na Esplanada dos Ministérios, expandiu-se e está presente em quase todas as regiões administrativas do Distrito Federal (TJDFT, 2020). Ciente dessa representatividade, em 2010, foi criado o Memorial Institucional, Espaço Desembargadora Lila Pimenta Duarte, como parte das comemorações dos 50 anos do tribunal, cadastrado no Instituto Brasileiro de Museus (IBRAM) com espaço físico e virtual.

Gonçalves *et al.* (2022) relatam que as atividades do Memorial estão amparadas por um programa destinado a divulgar a trajetória do tribunal desde a sua instalação, em Brasília, com um ambiente virtual e físico composto por documentos digitais históricos e artísticos, sob a tutela do Núcleo de Apoio à Preservação da Memória Institucional (NUAMI). Com o surgimento da pandemia causada pelo Coronavírus Disease 19 (COVID-19), as atividades presenciais foram suspensas temporariamente, com atenção às atividades virtuais e seus sistemas de informação.

O acervo do ambiente virtual consiste em obras artísticas e históricas, muitas das quais em fotografias. Com isso, geram-se muitas dúvidas sobre os direitos autorais, direitos morais e patrimoniais, além dos direitos de imagem, na medida em que se tem uma grande diversidade de contextos, como das imagens com uma ou várias personalidades, de fotografias de obras de arte, imagens digitais etc.

2. Objetivos

O presente trabalho tem como objetivo apresentar os resultados dos estudos voltados à análise dos direitos autorais dos documentos constantes no acervo digital do Memorial do TJDFT, apresentados no sistema de informação implementado com o software livre Omeka. Assim, busca explicitar as nuances que envolvem os direitos autorais, patrimoniais e de imagens de um acervo tão específico como o do Memorial do TJDFT.

3. Procedimentos Metodológicos

O estudo apresenta dados totalmente qualitativos, com análise na documentação mantida pelo Núcleo de Apoio à Preservação da Memória Institucional (NUAMI), responsável pelo Memorial do TJDFT, localizado na sede do tribunal em Brasília, a ser depositada em sistema de informação baseado no Omeka.

Assim, o estudo coletou dados na tipologia documental do acervo, incluindo todas as características que afetam os direitos autorais e de imagem. Posteriormente, analisou as informações com base na legislação relacionada aos direitos autorais.

4. Resultados

4.1 Sistema de Informação e Acervo Digital do Memorial do Tribunal de Justiça do Distrito Federal e dos Territórios

O Acervo do Memorial do TJDFT é amplo, abrangendo diversas tipologias documentais. Gonçalves *et al.* (2022) indica que esse tipo de acervo heterogêneo necessitaria de um sistema de informação próprio, visto que o seu potencial ficaria oculto se fosse incorporado ao acervo da biblioteca digital. Nesse sentido, utilizou-se o Omeka, por se tratar de um software livre e de código aberto.

Além dessas duas características, o Omeka possui outras mais específicas. Dentre elas, destacam-se a sua capacidade de interoperar com outros sistemas, por utilizar padrões internacionais de comunicação e descrição; a possibilidade de instalar plugins, fazendo com que as funcionalidades possam ser expandidas; o seu formato de disseminação,fazendo com que seja organizado em coleções e exposições, o que possibilita dar destaque para acervos inteiros ou para curadorias montadas a partir de vários; e a sua flexibilidade em utilizar diversos temas na

[1] https://www.tjdft.jus.br/informacoes/infancia-e-juventude/institucional/historia#:~:text=No%20ano%20de%201960%2C%20o,do%20Distrito%20Federal%20de%20Bras%C3%ADlia.

mesma instalação, oferecendo mais diversidade visual nas exposições.

Segundo exposto em Gonçalves (2022, p. 6), o Omeka corresponde ao software livre, desenvolvido pela *Roy Rosenzweig Center for History and New Media*, vinculado à *George Mason University*, atuando com o gerenciamento de coleções de objetos digitais. Segundo o autor:

> [...] o Omeka, que significa mostrar, espalhar, divulgar em Swahili, é um software pertencente aos conjuntos de ferramentas denominadas de GLAM (Galleries, Libraries, Archives and Museums), voltados à disseminação da informação, por meio de coleções de objetos digitais (GONÇALVES, 2022, p. 6).

Importante observar que o Omeka apresenta funcionalidades que permitem o uso flexível dos objetos digitais depositados, que podem assumir diferentes papéis, formatos e finalidades a depender do contexto, possibilitando diversidade de usos para único objeto digital. Nesse contexto, deve-se observar em relação aos objetos digitais a proteção atribuída aos direitos autorais.

4.2 Aspectos de Proteção de Direitos autorais

Os Direitos Autorais regulam os direitos de autor em sentido estrito e os direitos que lhes são conexos, protegendo os autores em relação às obras por eles criadas. O autor está conceituado em conformidade com o artigo 11 da LDA, compreendendo-se "[...] a pessoa física criadora de obra literária, artística ou científica" (BRASIL, 1998, on-line).

As obras intelectuais, objeto de proteção do direitos autorais, estão previstas no artigo 7º da LDA, ao serem definidas como "[...] as criações do espírito, expressas por qualquer meio ou fixadas em qualquer suporte, tangível ou intangível, conhecido ou que se invente no futuro" (BRASIL, 1998, on-line).

O artigo 22 da LDA estabelece que "pertencem ao autor os direitos morais e patrimoniais sobre a obra que criou" (BRASIL, 1998). Fundando-se em uma teoria dualista onde coexistem dois direitos basilares, quais sejam, de natureza moral e patrimonial.

No que se refere à transferência de direitos patrimoniais do autor, o artigo 49 da Lei de Direitos Autorais estabelece que os direitos de autor podem ser transferidos por meio de licenciamento, concessão, cessão ou por outros meios admitidos em Direito, e estabelece que podem ser transferidos de forma total ou parcial, a título universal ou singular, a terceiros, pelo próprio autor ou por seus sucessores. Conforme especificado no artigo 49 e seus incisos, a transmissão total compreende todos os direitos de autor, salvo os de natureza moral e os expressamente excluídos por lei. Além disso, elucida que a transmissão total e definitiva somente será admitida mediante contrato escrito:

> Art. 49. Os direitos de autor poderão ser total ou parcialmente transferidos a terceiros, por ele ou por seus sucessores, a título universal ou singular, pessoalmente ou por meio de representantes com poderes especiais, por meio de licenciamento, concessão, cessão ou por outros meios admitidos em Direito, obedecidas as seguintes limitações:
>
> I - a transmissão total compreende todos os direitos de autor, salvo os de natureza moral e os expressamente excluídos por lei;
>
> II - somente se admitirá transmissão total e definitiva dos direitos mediante estipulação contratual escrita;
>
> III - na hipótese de não haver estipulação contratual escrita, o prazo máximo será de cinco anos, (BRASIL, 1998, on-line).

Importa também a distinção estabelecida entre a proteção dos direitos autorais e do direito de imagem, considerando o objeto de proteção de cada instituto.

Segundo exposto em Sousa e Sabanai (2021, p.194), considerando que, embora os direitos morais do autor constituam direitos de personalidade, os direitos autorais protegem a criação de obras intelectuais. Já os direitos de imagem, enquanto direitos fundamentais, protegem a expressão de uso de imagem:

> [...] os direitos autorais tratam especificamente da proteção sobre a criação de obras intelectuais. Por

> intermédio desta, pode-se definir quem são os autores, detentores ou titulares das respectivas obras, seus direitos, bem como as possibilidades de uso livre sobre a obra e suas limitações. Já o direito à imagem, segundo Tartuce (2019), constitui direito fundamental inerente à dignidade da pessoa humana, que trata da proteção sobre a expressão de uso da imagem, reconhecida como parte do direito de imagem, a qual pode ser transferida a terceiros (SOUSA, SABANAI, 2020, p. 194).

Diante disso, observa-se que embora apresentem objetos de proteção diferentes, instrumentos específicos podem ser utilizados para transferência de Direitos patrimoniais. Normalmente os termos de cessão de uso de imagem e direitos autorais podem ser utilizados tanto para obras artísticas, científicas e literárias, como podem ser adequados, considerando o objeto de proteção, para os direitos de imagem.

4.3 Análise dos Direitos Autorais e de imagem da Tipologia Documental que compõe o Acervo do TJDFT

Foram analisados os direitos autorais e de imagens, de fotos de pessoas, prédios e imagens digitais diversas, que compõem o arquivo entre 1960 e 2000. Em relação às obras de arte, foram analisadas as adquiridas ao longo da história do Tribunal, independentemente do tipo de aquisição e de doações, sendo a última entre 2013 e 2019.

Das adquiridas por doação, considerando o período de 2013 a 2019, foram identificadas 87 obras de arte, bem como em relação a doações de livros, com o quantitativo de 33 livros, conforme identificadas no Quadro 1.

Tanto em relação às obras de arte quanto em aos livros doados ao TJDFT, ambos apresentam termos de doação, constando a identificação do doador, e as características do material doado.

Em relação às obras adquiridas pelo TJDFT, foram identificadas 20, as quais não apresentam termos de doação. Entretanto, não foi possível identificar a forma de aquisição delas, se ocorreu por meio de compra, concurso ou doação.

Ainda em relação às obras de arte adquiridas no decorrer da história, observou-se a existência de dados que identificam obras já com autores falecidos e o ano de primeira menção considerando o pertencimento da obra ao TJDFT.

Quadro 1 - Obras de arte doadas ao TJDFT no período de 2013 a 2019

Ano	Tipologia	Quantidade
2019	Obras de arte	12
	Livros	8
2018	Obras de arte	15
	Livros	7
2017	Obras de arte	14
	Livros	3
2016	Obras de arte	17
	Livros	3
2015	Obras de arte	17
	Livros	4
2014	Obras de arte	10
	Livros	6
2013	Obras de arte	2
	Livros	2

Fonte: Elaborado pelos autores

Da tipologia documental levantada, observaram-se também os direitos autorais em relação às obras de arte selecionadas por edital de concurso, contabilizando 16 imagens. Com vistas a integrar o acervo do TJDFT, foram selecionadas esculturas e pinturas para ocupação de lugar de destaque nos Fóruns do TJDFT, as quais passaram a integrar o acervo da instituição, com temas obrigatoriamente relacionados à justiça.

Nesse caso, as obras apresentaram termo de cessão de uso de obra de arte (uso de imagem), autorizando, em caráter exclusivo e isento de qualquer ônus, o uso das obras para divulgação pelo Tribunal, transferindo os direitos patrimoniais para escolha dos meios de reprodução pelo tribunal, meios de divulgação, formato, etc., tudo o que for necessário para que a reprodução seja efetivada pelo TJDFT.

Em relação aos direitos autorais, observaram-se também os requisitos da originalidade nos termos da lei de direitos autorais e convenções internacionais sobre o tema, além da exigência do Termo de Autorização de Uso e Cessão de Direitos Autorais para o trabalho apresentado, devidamente assinado pelo candidato, com firma reconhecida, e exigência do critério da

criatividade. O edital também deixa claro que os trabalhos submetidos, não aprovados e não retirados no prazo estipulado, passarão a compor o acervo do Tribunal, sem qualquer ônus para o TJDFT perante seu autor.

No que se refere às fotos, foram identificadas 14 fotos de pessoas e 15 de prédios. Quanto a elas, não foram localizados subsídios que dessem suporte ou acesso a documentação registrada, porém, o Memorial identificou que os direitos advindos da transmissão da produção intelectual (foto de pessoas e prédios) pertencem ao Tribunal. Entretanto, vale ressaltar que ainda há fotografias não catalogadas no Omeka. Nesse sentido, os números ainda são inquantificáveis.

5. Considerações Finais

No tocante à tipologia documental que compõe o acervo do Memorial Institucional, de responsabilidade do Núcleo de Apoio à Preservação da Memória Institucional (NUAMI), encontra-se depositada em sistema de informação baseado no Omeka. O referido sistema permite observar as formas de transferência de titularidade em relação aos direitos autorais e de imagens.

A importância do acervo está em oferecer informações sobre a memória institucional do TJDFT, ressaltando a atenção às atividades virtuais, com os seus sistemas de informação.

O estudo coletou dados na tipologia documental do acervo, tornando-se necessário tecer considerações sobre características que se relacionam aos direitos autorais e de imagem.

Quanto aos direitos autorais das obras de arte adquiridas pelo tribunal por meio de concurso, atendem aos procedimentos especificados para proteção dos direitos de autor e de imagem, considerando-os institutos específicos de proteção a direitos fundamentais e de personalidade, solicitados no próprio edital.

No que se refere às obras por doações, da mesma forma, por meio do termo de doação realizou-se a transferência das obras para o TJDFT.

No entanto, em relação às demais obras adquiridas pelo TJDFT, as informações fornecidas não deixam claro o instrumento formal utilizado para transferência de direitos autorais e de imagem, como ocorre nas fotos de pessoas e de prédios.

Chega-se ao entendimento de que existe a necessidade de um estudo mais aprofundado para aquisição de maiores subsídios sobre em que condições o TJDFT regulamentou a transferência de direitos autorais e de imagem de fotos de prédios e pessoas que compõem o acervo.

Diante disso, pode-se dizer que uma revisão na documentação por parte do Memorial permite elucidar questões relacionadas ao tema e, ao mesmo tempo, ressaltar a importância da manutenção do acervo para a memória institucional do TJDFT.

Referências

BRASIL. **Lei nº 9.610, de 19 de fevereiro de 1998**. Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências. Publicada no Diário Oficial da União em 20 fev. 1998. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9610.htm . Acesso em: 09 set. 2022,

GONÇALVES, Maison R. M.; SHINTAKU, Milton; PITANGA, Betânia Martins; ARRUDA, Aline Cristina Costa de; MENDONÇA, Talitha Selvati Nobre; MACÊDO, Diego José; PAIVA, Guilherme Guth de. **Guia do Usuário do**

Omeka para o Tribunal de Justiça do Distrito Federal e dos Territórios (TJDFT). Brasília: Ibict; TJDFT, 2022. 47 p. ISBN 978-65-89167-28-0. Disponível em: http://labcotec.ibict.br/omp/index.php/edcotec/catalog/book/119. Acesso em: 12 maio 2022.

SOUSA, R. P. M.; SABANAI, N. L. I. A Proteção da Pessoa com Surdez e a Política Autoral de Obra Audiovisual. BRITO, R. F. et. al. (org.). **Tradução para Libras Escrita**: relatos sobre o processo de tradução e implementação do SignWriting em um sistema de revistas científicas para surdos. São Carlos: Editora Scienza, 2021.

TJDFT. Tribunal de Justiça do Distrito Federal e Territórios - 60 anos. 2020.Disponível em: https://www.tjdft.jus.br/institucional/imprensa/noticias/2020/abril/tjdft-rumo-aos-60-anos. Acesso em 09 set. 2022.

HL7 FHIR BASEADO EM W3C PROV PARA ALCANÇAR A PROVENIÊNCIA DE DADOS EM SISTEMAS DE INFORMAÇÃO EM SAÚDE

HL7 FHIR BASED ON W3C PROV TO ACHIEVE DATA PROVENANCE IN HEALTH INFORMATION SYSTEMS

Márcio José Sembay[1], Douglas Dyllon Jeronimo de Macedo[2], Alexandre Augusto Gimenes Marquez Filho[3]

(1) Universidade Federal de Santa Catarina (UFSC), marcio.sembay@posgrad.ufsc.br
(2) Universidade Federal de Santa Catarina (UFSC), douglas.macedo@ufsc.br
(3) Universidade Federal de Santa Catarina (UFSC), alexandre.agmf@gmail.com

Resumo

A proveniência de dados e a interoperabilidade são requisitos fundamentais para os Sistemas de Informação em Saúde (SIS) para garantir não apenas o intercâmbio efetivo e eficiente dos dados de saúde, mas também, para definir fontes de dados confiáveis. Esforços de vários anos em padrões internacionais como *Health Level 7 (HL7)* e *Fast Healthcare Interoperability Resources (FHIR)* com base no modelo de proveniência de dados do *World Wide Web Consortium (W3C)*, mais conhecido como W3C PROV, confirmam esses requisitos. O objetivo desta pesquisa bibliográfica e exploratória, com abordagem qualitativa, é analisar e caracterizar o uso do padrão HL7 FHIR baseado em W3C PROV, possibilitando observar o alcance da proveniência de dados em SIS. As análises realizadas neste estudo basearam-se na literatura, possibilitando caracterizar a proveniência de dados no uso do HL7 FHIR com base no W3C PROV, além de destacar a importante intersecção desses elementos no cenário dos SIS. Assim, este estudo considera o HL7 FHIR baseado no W3C PROV, como um dos padrões internacionais que mais contribuem para alcançar a proveniência de dados em SIS.

Palavras-chave: Proveniência de Dados; W3C PROV; Sistemas de Informação em Saúde; Interoperabilidade; HL7 FHIR.

Abstract

Data provenance and interoperability are fundamental requirements for Health Information Systems (HIS) to ensure not only the effective and efficient exchange of health data, but also to define reliable data sources. Multi-year efforts at international standards such as Health Level 7 (HL7) and Fast Healthcare Interoperability Resources (FHIR) based on the World Wide Web Consortium (W3C) data provenance model, better known as the W3C PROV, confirm these requirements. The objective of this bibliographical and exploratory research with a qualitative approach is to analyze and characterize the use of the HL7 FHIR standard based on W3C PROV, making it possible to observe the reach of the data provenance in HIS. The analysis carried out in this study were based on the literature, making it possible to characterize the data provenance in the use of the HL7 FHIR based on the W3C PROV, in addition to highlighting the important intersection of these elements in the HIS scenario. Thus, this study considers the HL7 FHIR based on the W3C PROV, as one of the international standards that most contribute to achieving data provenance in HIS.

Keywords: Data Provenance; W3C PROV; Health Information Systems; Interoperability; HL7 FHIR.

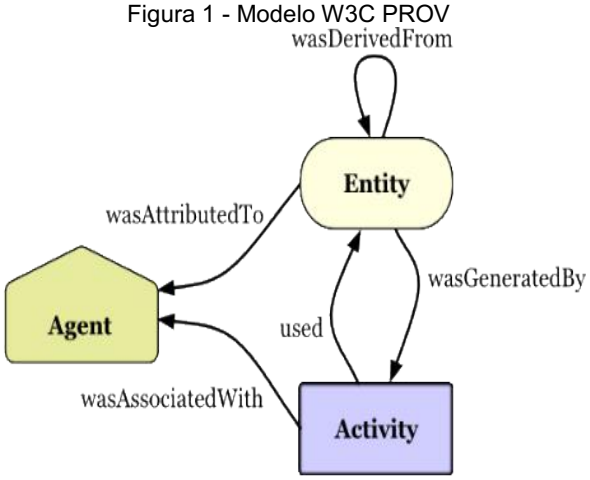
1. Introdução

Os Sistemas de Informação em Saúde (SIS) podem ser considerados como um Sistema de Informação (SI) que compõe a coleta, o processamento, a comunicação e o uso de informações fundamentais para aperfeiçoar a eficiência dos serviços de saúde (WORLD HEALTH ORGANIZATION, 2004). Sembay e Macedo (2022) ressaltam que os SIS têm a capacidade de gerar e armazenar grandes volumes de dados, tornando-se um desafio atual.

Nesse sentido, as abordagens da proveniência de dados contribuem nesse quesito para responder perguntas como: “por quê?”, “como?”, “onde?”, “quando?”, “por quem?” e “para quê?” os dados foram produzidos nos SIS, contribuindo para a rastreabilidade e gerenciamento dos dados de proveniência nos SIS (SEMBAY; MACEDO, 2022). A proveniência de dados identifica transformações pelas quais os dados passaram ao longo do tempo, e no contexto da saúde vive um cenário de pesquisa crescente, principalmente nas

tecnologias aplicadas nessa área que vêm obtendo resultados expressivos (SEMBAY; MACEDO; DUTRA, 2020a; SEMBAY; MACEDO; DUTRA, 2020b). No estudo de Freund, Sembay e Macedo (2019), os autores consideram que a proveniência de dados pode ser considerada como um requisito importante para estabelecer confiabilidade e prover segurança em SI. Dessa forma, na Figura 1, destacamos o modelo de proveniência de dados PROV criado em 2013. (GIL; MILES, 2013).

Figura 1 - Modelo W3C PROV



Fonte: Gil e Miles (2013).

Conforme a Figura 1, existem três tipos diferentes de elementos no W3C PROV (GIL; MILES, 2013): *Entity* (entidades), *Activity* (atividades) e *Agent* (agente). Uma entidade é um tipo físico, digital, conceitual ou outro tipo de coisa com alguns aspectos fixos. Uma atividade é algo que age sobre ou com entidades; pode incluir consumir, processar, transformar, modificar, realocar, usar ou gerar entidades. Finalmente, um agente é algo que carrega alguma forma de responsabilidade por uma atividade que ocorre, pela existência de uma entidade ou pela atividade de outro agente. Entre as entidades, atividades e agentes, pode haver diferentes tipos de relações como apresentado na Figura 1.

Nesse sentido, é importante ressaltar que o W3C PROV aplicado em SIS, pode contribuir para que os dados de saúde possam ser interoperáveis. Assim, destacamos a interoperabilidade de software no domínio da saúde utilizando um padrão consistente como *Health Level Seven (HL7)* e *Fast Healthcare Interoperability Resources (FHIR)*. O HL7 é uma organização de desenvolvimento de padrões para troca, integração, compartilhamento e recuperação de informações de saúde. O FHIR, criado pelo HL7, é um padrão que descreve formatos e elementos de dados para troca de registros eletrônicos de saúde interoperáveis (HL7 FHIR, 2022a). Porém, Landgrebe e Smith (2011) afirmam que o HL7 FHIR pode ter diversos desafios de interoperabilidade. Nesse sentido, Margueri *et al* (2020) ressalta as bases de proveniência de recurso do HL7 FHIR com o W3C PROV, possibilitando além de compartilhar atributos de proveniência, sanar problemas de interoperabilidade semântica.

Assim, para que todas as relações de proveniência de dados do HL7 FHIR baseadas no W3C PROV sejam realizadas, é importante destacar o padrão HL7 FHIR 5Ws que inclui as seguintes questões: *Who?* (Quem?), *What?* (O quê?), *When?* (Quando?), *Where?* (Onde?), *Why?* (Por quê?). Daí o nome deste padrão (5Ws) (HL7 FHIR, 2022b). Esse padrão também inclui muitos recursos, possibilitando mapear a proveniência com base no W3C PROV (HL7 FHIR, 2022c).

Portanto, este estudo tenta responder a seguinte questão: Qual a importância do uso do padrão HL7 FHIR com base no W3C PROV para alcançar a proveniência de dados em cenários de SIS?

2. Objetivo

Este estudo tem como objetivo analisar e caracterizar o uso do padrão HL7 FHIR baseado em W3C PROV, possibilitando observar o alcance da proveniência de dados em SIS.

3. Procedimentos Metodológicos

No presente estudo, realizamos uma pesquisa de caráter bibliográfico e exploratório de natureza qualitativa acerca da temática do uso do HL7 FHIR, com base no W3C PROV para alcançar a proveniência de dados em SIS. Adiante, os métodos deste estudo são descritos em três passos para a realização das análises.

O primeiro passo é caracterizar a proveniência de dados em relação ao HL7 FHIR com base no W3C PROV. Para isso, comparamos as questões base do padrão

HL7 FHIR 5Ws (HL7 FHIR, 2022b) responsável pelo mapeamento dos recursos de proveniência em SIS, juntamente com as questões base da proveniência de dados definidas por Buneman, Khanna e Wang-Chiew (2001). Logo depois, são caracterizadas as aplicações da proveniência de dados de parte da taxonomia definida por Simmhan *et al.* (2005): i) qualidade dos dados (baseado nos dados originais e nas suas derivações); ii) trilha de auditoria (detectando erros na geração de dados); iii) receitas de replicação (repetição da derivação de dados e sua atualização); iv) atribuição (direitos autorais e propriedade de dados); e, v) informativo (interpretação de dados). Essas aplicações são caracterizadas com o HL7 FHIR baseado no W3C PROV de acordo com Margheri *et al.* (2020) e Kohlbacher *et al.* (2018), destacando sua importância de aplicação em SIS.

O segundo passo é demonstrar por meio da literatura científica a importância do uso do HL7 FHIR baseado no W3C PROV em SIS. Para validar esse passo, selecionamos cinco artigos na literatura científica de pesquisas recentes que envolvessem soluções dirigidas para alcançar a proveniência de dados em SIS por meio do HL7 FHIR baseado em W3C PROV, observando as importantes visões consolidadas nas pesquisas nesta área. Assim, deste corpus científico foi possível realizar as análises para apresentar os principais termos e tecnologias identificadas no uso do HL7 FHIR baseado no W3C PROV em diferentes tipos de SIS.

Por fim, no terceiro passo, apresentamos a proveniência de dados como intersecção entre o HL7 FHIR com base no W3C PROV e HL7 FHIR 5Ws, ambos como fatores de importância e necessidade em relação à interoperabilidade existente nos cenários dos SIS.

4. Análises e Resultados

As análises realizadas basearam-se na metodologia descrita neste estudo e seguem descritas nas próximas subseções.

4.1 Caracterizando a proveniência de dados em HL7 FHIR com base no W3C PROV para SIS

A Tabela 1 apresenta as questões do padrão HL7 FHIR 5Ws definidos pelo HL7 FHIR (2022b) com as questões de proveniência de dados definidas por Buneman, Khanna e Wang-Chiew (2001).

Tabela 1 - Comparando as questões base - HL7 FHIR 5Ws com a proveniência de dados em SIS

Padrão HL7 FHIR 5Ws	**Proveniência de Dados**
Who? (Quem?)	Who? (Quem?)
What? (O que?)	How? (Como?)
When? (Quando?)	When? (Quando?)
Where? (Onde?)	Where? (Onde?)
Why? (Por quê?)	Why? (Por quê?)

Fonte: dados da pesquisa (2022).

Pode-se observar na Tabela 1 que praticamente todas as questões comparadas são iguais, confirmando a necessidade da existência da proveniência de dados para que haja interoperabilidade de dados realizada no HL7 FHIR com base no W3C PROV, que de fato, para o domínio dos SIS é significativo. Outro ponto importante observado na Tabela 1, é que o padrão HL7 FHIR 5Ws possibilita o mapeamento de recursos de proveniência por se basear nas premissas do W3C PROV e consequentemente contribui para a segurança usada no rastreamento, status, histórico e autoria de dados para diferentes SIS. Já na Tabela 2 é apresentada a caracterização das aplicações da proveniência de dados definidas por Simmhan *et al.* (2005) em relação ao HL7 FHIR baseado no W3C PROV conforme indicado por Margheri *et al.* (2020) e Kohlbacher *et al.* (2018).

Tabela 2 - Caracterizando aplicações da proveniência de dados com HL7 FHIR baseado no W3C PROV para SIS

Aplicações da Proveniência de Dados	**HL7 FHIR baseado em W3C PROV**
Qualidade dos Dados	Estima a qualidade e a confiabilidade dos dados com base nas transformações de origem e dados.
Trilha de	Determina o uso de recursos e

Auditoria	detecta erros nos dados de geração. Evento de Auditoria de Recursos.
Receitas de replicação	Permite a replicação e derivação de dados através de informações detalhadas de procedência.
Atribuição	Isso possibilita determinar a propriedade e os direitos sobre os dados.
Informativo	O *Representational State Transfer (RESTful)*, uma arquitetura regrada de software, mantém o controle das informações de proveniência relacionadas a recursos.

Fonte: dados da pesquisa 2022.

A Tabela 2 caracteriza fatores que contribuem para os SIS nos seguintes aspectos: i) gestão de dados para profissionais de saúde; ii) controle social dos SIS, favorecendo a coleta qualificada e a interoperabilidade de informações (dados sensíveis) que apoiam o acesso à saúde, bem como a alocação de recursos de proveniência; e, iii) possibilitando a interoperabilidade de dados nos SIS, que de fato é um dos maiores desafios do setor.

Ainda, enfatizando a Tabela 2, no estudo de Sembay, Macedo e Marquez Filho (2022), os autores comparam as mesmas aplicações da proveniência de dados com a tecnologia Blockchain, também responsável por contribuir com a proveniência de dados em SIS.

4.2 Usando HL7 FHIR baseado no W3C PROV em SIS

Com base na metodologia descrita para esta pesquisa, destacamos cinco estudos encontrados na literatura, os quais possibilitaram extrair oito principais termos apresentados na Tabela 3.

Tabela 3 - Principais termos que caracterizam o uso do HL7 FHIR baseado no W3C PROV em SIS

Autor/Ano	**Termos**	**HL7 FHIR baseado em W3C PROV**
Bittins *et al.* (2021), Margheri *et al.* (2020), Massi *et al.* (2018), Kohlbacher *et al.* (2018), Mense e Blobel (2017)	Interoperabilidade	Sintática e semântica
	Segurança	*Hyper Text Transfer Protocol Secure (HTTPS)*, - *Secure Sockets Layer (SSL)*
Bittins *et al.* (2021) Massi *et al.* (2018), Margheri *et al.* (2020)	Criptografia	Além dos protocolos HL7 v2, ele suporta outras normas
Bittins *et al.* (2021) Margheri *et al.* (2020), Massi *et al.* (2018), Kohlbacher *et al.* (2018), Mense e Blobel (2017)	Privacidade	Implementar privacidade e fácil devido às *Application Programming Interface (APIs)*
Bittins *et al.* (2021) Kohlbacher *et al.* (2018)	Compatibilidade	Facilmente integrado com: *Digital Imaging and Communications in Medicine (DICOM)*
Massi *et al.* (2018), Kohlbacher *et al.* (2018)	Flexibilidade	Muito flexível
Bittins *et al.* (2021) Margheri *et al.* (2020), Massi *et al.* (2018), Kohlbacher *et al.* (2018), Mense e Blobel (2017)	Proveniência	Modelo W3C PROV
	Arquitetura	Com base no protocolo *RESTful*

Fonte: dados da pesquisa 2022.

Em relação aos oito termos apresentados na Tabela 3, podemos destacar que todos são discutidos no W3C PROV para que o processo de proveniência seja possível através do padrão HL7 FHIR. Indo além, a Tabela 4 apresenta três tipos de SIS: (*Electronic Health Record (EHR)*, *Personal Health Record (PHR)* e *Electronic Medical Record (EMR)* e as principais tecnologias utilizadas em conjunto de aplicações do HL7 FHIR com base no W3C PROV, sendo elas: Blockchain, *Integrating the Healthcare Enterprise (IHE)*, DICOM e *Clinical Document Architecture* (CDA).

Tabela 4 - Tipos de SIS e tecnologias que contribuem para o uso do HL7 FHIR baseado no W3C PROV em SIS

Autor/Ano	Tipos de SIS	Principais tecnologias
Bittins *et al.* (2021)	EHR	HL7 FHIR, IHE, DICOM, CDA e Blockchain
Margheri *et al.* (2020)	EHR	HL7 FHIR, Blockchain e IHE
Massi *et al.* (2018)	EHR	HL7 FHIR, Blockchain e IHE
Kohlbacher *et al.* (2018)	EMR	HL7 FHIR, IHE, DICOM e CDA
Mense e Blobel (2017)	EHR e PHR	HL7 FHIR e CDA

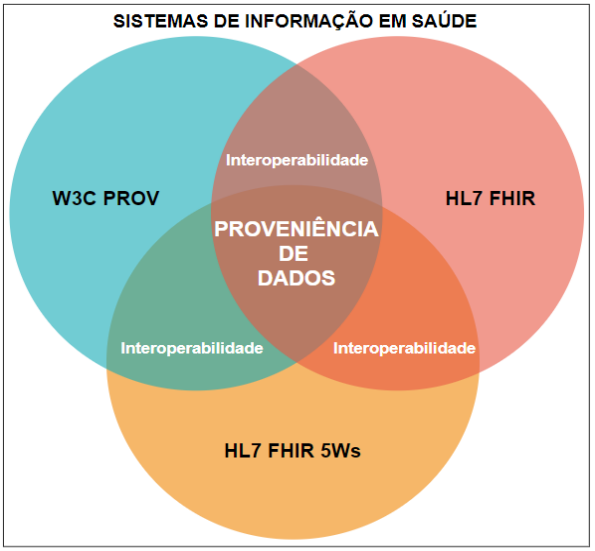
Fonte: dados da pesquisa (2022).

Observa-se na Tabela 4 que o EHR se destaca nos estudos de Bittins *et al.* (2021), Margheri *et al.* (2020), Massi *et al.* (2018) e Mense e Blobel (2017), por se tratar do SIS mais popular e mais utilizado em vários países. Logo após, o EMR que garante o registro médico protegido apresentado no estudo de Kohlbacher *et al.* (2018) e o PHR, relacionado a dispositivos móveis, apresentado no estudo de Mense e Blobel (2017), são considerados essenciais pelos autores. É importante ressaltar que, tanto no EHR, PHR e EMR os dados podem ser armazenados de forma distribuída, possibilitando a proveniência de dados em saúde. A Tabela 4 também destaca o uso das seguintes tecnologias que auxiliam no alcance da proveniência de dados em SIS: Blockchain nos estudos de Bittins *et al.* (2021), Margheri et al (2020) e Massi et al (2018) no quesito de interoperabilidade e imutabilidade dos dados; IHE nos estudos de Bittins *et al.* 2021), Margheri *et al.* (2020), Massi *et al.* (2018) e Kohlbacher *et al.* (2018) permitindo a melhoria da interoperabilidade; e por fim, CDA encontrado nos estudos de Bittins *et al.* (2021), Kohlbacher *et al.* (2018) e Mense e Blobel (2017) e DICOM encontrados nos estudos de Bittins *et al.* (2021) e Kohlbacher *et al.* (2018) relacionados à alta sensibilidade dos dados de saúde.

4.3 Proveniência de dados como intersecção entre o HL7 FHIR com base no W3C PROV e HL7 FHIR 5Ws no cenário de SIS

A Figura 2 apresenta a proveniência de dados como intersecção entre HL7 FHIR/ W3C PROV/HL7 FHIR 5Ws para mapear a proveniência de recursos nos cenários de SIS.

Figura 2 - Proveniência de dados como intersecção entre HL7 FHIR/W3C PROV/HL7 FHIR 5Ws



Fonte: Elaborada pelo autor.

Na Figura 2, a proveniência de dados, mostra-se como o principal insumo e benefício de trabalho do HL7 FHIR baseado no W3C PROV, oportunizando a geração de conhecimento e contribuindo para a tomada de decisão concisa nos setores de saúde.

5. Considerações Finais

Os resultados obtidos das análises realizadas vão ao encontro do objetivo proposto nesta pesquisa e têm o potencial de contribuir com futuros estudos que visem melhorar as práticas associadas ao uso do HL7 FHIR baseado em W3C PROV, que apesar de enfrentar desafios de interoperabilidade, muitas vezes únicos para cada instituição, continua sendo aperfeiçoado para contribuir no alcance da proveniência de dados em SIS. A importância do uso do padrão HL7 FHIR com base no W3C PROV foi evidenciada neste recente e dinâmico campo de estudos que se forma na intersecção da proveniência de dados no cenário de SIS. Considera-se ainda que as

contribuições identificadas neste estudo são representativas do potencial impacto exercido nas práticas científicas da Ciência da Computação e da Ciência da Informação, prospectando trabalhos futuros que visem ampliar as reflexões e contribuições aqui descritas.

Referências

BITTINS, Soeren *et al.* Healthcare data management by using blockchain technology. *In*: **Applications of blockchain in healthcare**. Singapore: Springer, 2021. p. 1-27.

BUNEMAN, Peter; KHANNA, Sanjeev; WANG-CHIEW, Tan. **Why and where**: a characterization of data provenance. *In*: INTERNATIONAL CONFERENCE ON DATABASE THEORY. Berlin, Heidelberg: Springer, 2001. p. 316-330.

FREUND, Gislaine P.; SEMBAY, Márcio J.; MACEDO, Douglas D. J. de. Proveniência de Dados e Segurança da Informação: relações interdisciplinares no domínio da ciência da informação. **Revista Ibero-Americana de Ciência da Informação**, [*s. l.*], v. 12, n. 3, p. 807-825, 14 set. 2019. Disponível em: https://periodicos.unb.br/index.php/RICI/article/view/21203/23548. Acesso em: 20 maio 2022.

GIL, Yolanda; MILES, Simon. **PROV Model Primer**. *In*: W3C, 2013 April 30. W3C Working Group Note. Disponível em: https://www.w3.org/TR/prov-primer/. Acesso em: 02 ago. 2022.

HL7 FHIR. **Welcome to FHIR**. 2022a. Disponível em: Disponível em: https://hl7.org/fhir/. Acesso em: 02 ago. 2022.

HL7 FHIR. **5Ws Pattern**. Pattern FiveWs: Content. 2022b. Disponível em: https://build.fhir.org/fivews.html. Acesso em: 02 ago. 2022.

HL7 FHIR. **Provenance**. 2022c. Disponível em: https https://www.hl7.org/fhir/provenance.html. Acesso em: 05 ago. 2022.

KOHLBACHER, Oliver *et al.* **Data Integration for Future Medicine (DIFUTURE)**: an architectural and methodological overview. 2018. DOI https://doi.org/10.3414/ME17-02-0022

LANDGREBE, Jobst; SMITH, Barry. The HL7 approach to semantic interoperability. *In:* **Proceedings of the 2nd International Conference on Biomedical Ontology.** CEUR, vol. 833. pp. 139-146, 2011.

MARGHERI, Andrea *et al.* Decentralised provenance for healthcare data. **International Journal of Medical Informatics**, v. 141, p. 1-21, set. 2020. Elsevier BV. DOI http://dx.doi.org/10.1016/j.ijmedinf.2020.104197.

MASSI, Massimiliano *et al.* **Using PROV and blockchain to achieve health data provenance**. 2018. Disponível em: https://eprints.soton.ac.uk/421292/1/PROV_BC_Healthcare.pdf. Acesso em: 02 ago. 2022.

MENSE, Alexander; BLOBEL, Bernd. HL7 standards and components to support implementation of the European general data protection regulation. **European Journal for Biomedical Informatics**, v. 13, n. 1, p. 27-33, 2017.

SEMBAY, Márcio José; MACEDO, Douglas Dyllon Jeronimo de. Sistemas de informação em saúde: proposta de um método de gerenciamento de dados de proveniência no instanciamento do modelo W3C PROV-DM. **Advanced Notes in Information Science**, v. 2, p. 192-201,Tallinn, Estonia: ColNes Publishing, 2022. DOI: 10.47909/anis.978-9916-9760-3-6.101.

SEMBAY, Márcio José; MACEDO, Douglas Dyllon Jeronimo de; MARQUEZ FILHO, Alexandre Augusto Gimenes. Identification of the relationships between Data Provenance and Blockchain as a contributing factor for Health Information Systems. *In:* EAI INTERNATIONAL CONFERENCE ON DATA AND INFORMATION IN ONLINE ENVIRONMENTS, 3., 2022, Florianópolis. ***Proceedings*** [...],. Florianópolis: DIONE 2022, p. 189-203.

SEMBAY, Márcio José; MACEDO, Douglas Dyllon Jeronimo de; DUTRA, Moisés Lima. L. A proposed approach for provenance data gathering. **Mobile Networks & Applications,** p. 1-13, 2020a. DOI 10.1007/s11036-020-01648-7

SEMBAY, Márcio José; MACEDO, Douglas Dyllon Jeronimo de; DUTRA, Moisés Lima. A Method for collecting provenance data: a case study in a Brazilian hemotherapy center. *In*: **Lecture Notes of the Institute for Computer Sciences, Social InformaTIC and Telecommunications Engineering**. 1. ed. [*S. l.*]: Springer International Publishing, 2020b. v. 1, p. 89-102. DOI 10.1007/978-3-030-50072-6_8

SIMMHAN, Yogesh L. *et al.* A survey of data provenance techniques. **Computer Science Department**, Indiana University, Bloomington IN, v. 47405, p. 69, 2005.

WORLD HEALTH ORGANIZATION *et al*. **Developing health management information systems**: a practical guide for developing countries. 2004.

IMAGO: UMA PROPOSTA PARA O BANCO DE IMAGENS DO INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA

IMAGO: A PROPOSAL FOR THE IMAGE BANK OF THE BRAZILIAN INSTITUTE OF INFORMATION IN SCIENCE AND TECHNOLOGY

Diego José Macedo[1], Ítalo Barbosa Brasileiro[2], Milton Shintaku[3]
(1) Instituto Brasileiro de Informação em Ciência e Tecnologia, Brasília - DF, diegomacedo@ibict.br
(2) Instituto Brasileiro de Informação em Ciência e Tecnologia, Brasília - DF, italobrasileiro@ibict.br
(3) Instituto Brasileiro de Informação em Ciência e Tecnologia, Brasília - DF, shintaku@ibict.br

Resumo
A adoção de imagens para representação de informação tem grande afinidade com a ciência e a tecnologia, na medida em que pode ser usada para diversos propósitos no processo de geração do conhecimento. No entanto, há desafios em gerenciar de forma sistêmica esses recursos informacionais considerando o processo de seleção, análise, guarda e reúso. Nesse sentido, o presente estudo tem o objetivo de estruturar um banco de imagens em ciência e tecnologia do Instituto Brasileiro de Informação em Ciência e Tecnologia, considerando aspectos de prospecção e implantação de tecnologia e definição de orientações para o gerenciamento do banco. Metodologicamente, foram definidos 12 critérios para seleção e análise do *software* e desenvolvidas orientações de acordo com os estudos apresentados. Como resultado, foram analisados 10 softwares de gerenciamento de imagens, e o software Piwigo foi selecionado por atender todos os critérios definidos. Além do processo de seleção do software, este estudo apresenta as normas de utilização estabelecidas para o uso dos recursos imagéticos que compõem o banco de imagens do instituto. A avaliação dos critérios e a definição dessas diretrizes para a utilização do banco de imagens garantem o fluxo adequado para a implantação da ferramenta e utilização pública dos recursos.

Palavras-chave: ciência e tecnologia; banco de imagens; critérios de avaliação; diretrizes de utilização.

Abstract
The adoption of images to information representation has great affinity with science and technology, as they can be used for different purposes in the process of knowledge generation. However, there are challenges in how to systematically manage these information resources considering the selection, analysis, storage and reuse process. In this sense, the present study has the objective of structuring an image bank in science and technology for the Brazilian Institute of Information in Science and Technology, considering aspects of prospection and implementation of technology and definition of guidelines for the bank's management. Methodologically, 12 criteria were defined for the selection and analysis of the software and guidelines were developed according to the studies presented. As a result, 10 image management software were analyzed, and the Piwigo software was selected for meeting all the defined criteria. In addition to the software selection process, this study presents the usage rules established for the use of the image resources that compose the institute image bank. The evaluation of the criteria and the definition of these guidelines for the use of the image bank guarantee the adequate flow for the implementation of the selected tool and the public use of the resources.

Keywords: science and technology; image bank; evaluation criteria; usage guidelines.

1 Introdução

A Ciência e Tecnologia (C&T) e as imagens têm um relacionamento antigo, na medida em que as imagens têm capacidade de representar informações científicas e tecnológicas, pois independentemente do tipo ou formato, são instrumentos úteis à C&T, na função de apoio ao relato científico. A importância do amparo teórico sobre a informação imagética na ciência da informação é relatada em Figueiredo e Saldanha (2017), baseado em Paul Otlet. Em muitas pesquisas, grande quantidade de imagens são criadas, de grande valor à ciência e à tecnologia. Entretanto, essas imagens quase sempre apresentam acesso restrito, dificultando a sua utilização e reprodução nas publicações. Assim, a manutenção dos recursos gráficos em um banco de imagens proporciona uma solução para armazenamento, disseminação, compartilhamento e reúso, alinhado às premissas da Ciência Aberta.

Além disso, a construção de um banco de imagens apresenta alguns desafios, como a seleção de tecnologia a ser utilizada, as questões relacionadas aos direitos autorais, a utilização de licenças associadas às imagens do banco etc.

Nesse contexto, o presente estudo tem por objetivo relatar o processo de seleção da ferramenta a ser utilizada no projeto e implementação de um banco de imagem em ciência e tecnologia, denominado "Imago", desenvolvido pelo Ibict. O Imago tem como função a promoção de competências e desenvolvimento de recursos de informação em ciência e tecnologia.

O desenvolvimento desse estudo de prospecção e implantação do banco de imagens fornece ferramentas para atender duas grandes questões relacionadas. A primeira delas é a escolha da ferramenta adequada para formar o banco de imagens, e a segunda é a definição de orientações para a utilização do banco e reprodução dos itens do acervo. Neste artigo são apresentados estudos desenvolvidos para atender os dois pontos mencionados.

2 Objetivos

O objetivo deste trabalho é fornecer a visão geral das ferramentas disponíveis para criação e manutenção de bancos de imagens, além de discutir algumas diretrizes para acesso e utilização de bancos de imagens de domínio público.

O primeiro objetivo principal consiste em analisar as tecnologias livres para implantação do banco de imagens, considerando um conjunto de critérios levantados.

O segundo objetivo é definir um conjunto de normas que devem ser atendidas para a utilização do banco de imagens. As orientações devem ser elaboradas de forma a não causar, para as instituições que mantêm a base, algum prejuízo pela má utilização dos recursos imagéticos por terceiros.

3 Procedimentos Metodológicos

Na primeira fase do estudo, foi realizado o levantamento de um conjunto de critérios para avaliação das tecnologias voltadas para criação de banco de imagens do instituto.

Os critérios selecionados foram observados entre os requisitos levantados em algumas referências, além de atender às orientações do Ibict. Foram avaliados trabalhos como o de Hayati *et al.* (2020), Hoitink (2007), Rahi, Ghani e Ngah (2020), Rodrigues (2007), entre outros, e escolhidos critérios que melhor se amoldam ao tipo de ferramenta e à finalidade pretendida.

O conjunto de critérios foi avaliado por um pequeno grupo de pesquisadores, envolvidos no processo de escolha e implantação do banco de imagens. Os critérios estabelecidos para a escolha da ferramenta são os seguintes:

A. Ferramenta open-source - como o objeto de manutenção do software é uma instituição pública, a ferramenta ideal deve ser gratuita e de código aberto.

B. Suporte para grande volume de imagens - por se tratar de um banco de imagens voltado à pesquisa e, além disso, para manter a memória institucional, o software deve suportar grande volume de dados, sem comprometer seu desempenho e a disponibilidade dos recursos.

C. Acessibilidade por interface web - para que a mesma base de imagens seja acessada por grande quantidade de usuários em diferentes localidades e horários, é preciso garantir um acesso aos dados em nuvem. Para isso, o acesso à ferramenta deve se dar por navegador web com conexão à Internet.

D. Criação, edição e remoção de imagens em lote - o software deve permitir a inserção e modificação de múltiplas imagens simultaneamente, garantindo agilidade no gerenciamento do banco.

E. Suporte para as principais extensões de imagens - o software deve ser capaz de armazenar e manipular diferentes extensões de imagens, especialmente os formatos mais comuns de arquivos.

F. Ferramenta extensível - a ferramenta deve apresentar código alterável e capacidade de extensão. Assim, o banco de imagens apresentará maior flexibilidade para a criação e adaptação de novas funcionalidades.

G. Gerenciamento de usuários e grupos de usuários - o software escolhido deve viabilizar a criação e o gerenciamento de usuários e diferentes grupos de usuários,

considerando seus diferentes papéis e a hierarquia dentro do sistema.

H. Controle de permissões - os usuários e grupos de usuários devem possuir restrições de acesso, de acordo com seu papel dentro do sistema. A permissão deve ser definida pelo(s) administrador(es) do sistema. Da mesma forma, deve ser possível definir quais usuários terão papel de administradores.

I. Manutenção dos metadados - os metadados descritivos das imagens do banco devem ser mantidos, tendo em vista que os mesmos podem ser objetos de pesquisa.

J. Criação de hierarquia entre os álbuns - deve ser garantida a capacidade de separar as imagens em diferentes álbuns. Além disso, também deve ser permitido organizar os álbuns em níveis de hierarquia e criar associações de pertencimento entre eles.

K. Multiplataforma - o software deverá ser acessado em diferentes aparelhos e sistemas operacionais, sem a ocorrência de perda de desempenho.

L. Documentação e comunidade ativa e participativa - deve haver conteúdo de documentação para a ferramenta e um suporte à comunidade, pois as duas características são indispensáveis à solução de eventuais problemas.

Para a criação dos critérios e orientações de utilização dos recursos imagéticos, foi feito o levantamento de regras criadas para outros bancos de recursos. A principal fonte foi a política de uso de imagens do banco de imagens estabelecidos em Ibama (2022).

A seguir, é feita a descrição das ferramentas disponíveis para banco de imagens. Depois é feita a avaliação dos softwares, de acordo com os requisitos levantados. Por fim, são apresentadas as orientações de utilização de imagens públicas que compõem o banco de imagens do Imago.

4 Resultados

O método utilizado para seleção de tecnologias foi a verificação, na web, dos softwares que compreendem a função principal de gerenciar acervos de imagens. Desse modo, para análise de tecnologias, foram encontradas as seguintes ferramentas:

- **Piwigo**: software de manutenção de fotos de código aberto, que permite o gerenciamento, a organização e o compartilhamento de fotos na web, projetado para organizações, equipes e indivíduos;
- **Nextcloud**: software que facilita a sincronização, o compartilhamento e a colaboração em seus arquivos;
- **ImageGlass**: visualizador de imagens básico de código aberto, que, embora simples, se beneficia da velocidade de ser leve e uma boa opção para usuários do Windows.
- **Gwenview**: visualizador de imagens rápido e fácil de usar do KDE, ideal para navegar e exibir uma coleção de imagens.
- **LightZone**: programa de edição de fotos que permite manipular imagens da mesma maneira que seria possível em um laboratório de fotos tradicional.
- **Darktable**: aplicativo de fluxo de trabalho de fotografia, de código aberto e desenvolvedor bruto, que gerencia negativos digitais em um banco de dados e permite visualizá-los por meio de uma mesa de luz com zoom, tornando possível desenvolver imagens brutas e aprimorá-las.
- **jAlbum**: aplicativo que permite criar álbuns de fotos on-line profissionais para qualquer site.
- **PhotoPrism**: software com inteligência artificial para navegar, organizar e compartilhar uma coleção de fotos mediante uso de tecnologias para marcar e encontrar fotos automaticamente, sem atrapalhar.
- **Lychee**: ferramenta gratuita de gerenciamento de fotos que roda servidor, permitindo carregar, gerenciar e compartilhar fotos como de um aplicativo nativo.
- **Jellyfin**: solução de mídia criada para controle de mídias, possibilitando a transmissão para qualquer dispositivo a partir de servidor, sem restrições.

Abaixo, é apresentado um quadro comparativo, com a avaliação dos critérios listados na seção três das ferramentas de banco de imagens elencadas acima. Os critérios são indicados pelas letras utilizadas em seu respectivo tópico.

Quadro 01 - Quadro comparativo entre os softwares

	Requisitos											
Software	A	B	C	D	E	F	G	H	I	J	K	L
Piwigo	x	x	x	x	x	x	x	x	x	x	x	x
Nextcloud	x	x	x	-	x	-	x	x	x	-	x	x
ImageGlass	x	x	-	-	x	-	-	-	x	-	-	-
Gwenview	-	x	-	-	x	-	-	-	x	x	-	x
LightZone	-	-	-	-	x	-	-	-	-	-	-	-
darktable	x	-	-	-	x	x	-	-	x	-	-	x
jAlbum	-	x	x	-	x	-	-	-	x	x	-	x
PhotoPrism	x	x	x	-	x	x	-	-	x	-	x	x
Lychee	x	x	x	x	x	x	-	-	x	-	x	x
Jellyfin	x	x	x	-	-	-	-	-	-	-	x	x

Fonte: Elaboração dos autores (2022).

O levantamento das tecnologias para a construção do Imago, Banco de Imagem em Ciência e Tecnologia, a partir da prospecção das tecnologias disponíveis, possibilitou a avaliação de softwares conforme os requisitos estabelecidos, sendo o Piwigo o único que atendeu a todos eles.

Com base em uma análise aprofundada do Piwigo, que levou em conta a instalação e os testes no software, foi possível confirmar os critérios e suas funcionalidades. Além disso, a exploração de suas principais funções constatou que a ferramenta cumpre com seus objetivos.

O seguinte conjunto de orientações foi concebido para definir e esclarecer algumas práticas para o gerenciamento do Imago, além de compreender informações de licença de acesso e uso das imagens que compõem o banco.

Para melhor estruturar as regras, o conjunto de orientações foi dividido em três grupos, que estão descritos a seguir.

- **Sobre os serviços**

1.1 Os serviços disponibilizados e a permissão de uso dos itens do acervo inseridos no Imago seguem os termos e condições das legislações e normas técnicas aplicáveis.
1.2. O acesso aos recursos do Imago se dará por meio de interface web, disponível a qualquer usuário que tenha Internet. Portanto, para se conectar ao banco de imagens é indispensável o uso de equipamento com requisitos mínimos de software e hardware. A equipe administrativa do Imago não se responsabiliza por possíveis problemas de acesso por falhas de conexão à Internet ou por utilização de software ou hardware de configuração inferior aos requisitos mínimos.
1.3. As imagens constantes no acervo estão sujeitas à legislação brasileira de direitos autorais. O banco de imagens pode ser utilizado livremente, sem custo ou autorização associada, acordando com a licença Creative Commons. Esse modelo de licença viabiliza o incentivo e a disseminação da informação tecnológica e científica vinculada ao Ibict. Por fim, é exigida a atribuição dos créditos das imagens, de acordo com o modelo Nome do autor/Ibict, para manutenção do direito moral do autor.

- **Sobre o acervo**

2.1. A equipe administrativa do Imago tem total liberdade e direito para remover ou editar qualquer registro contido no banco de Imagens, em qualquer momento.
2.2. Excetuando em casos de manutenção de sistema, o conteúdo do Imago estará disponível on-line, em tempo integral, e de forma gratuita para o acesso de usuários.
2.3. O conteúdo do acervo do Imago é de propriedade exclusiva do Ibict e dos seus autores, de acordo com as normas de Propriedade Intelectual e Direitos Autorais.

- **Sobre as responsabilidades**

3.1. O Ibict não tem responsabilidade por uso indevido das imagens do acervo, cabendo ao usuário toda a responsabilidade por quaisquer violações. Caso o usuário não garanta a veracidade e a exatidão das informações prestadas, poderão ser tomadas medidas legais cabíveis pelo uso indevido.
3.2. A utilização de alguma imagem pertencente ao acervo do Imago obriga o usuário a citar, de forma clara e legível, os créditos do autor e da fonte, no formato Nome do autor/Banco de Imagens do Ibict. A omissão dos créditos representa violação do direito autoral e pode gerar penalidades previstas em legislação.
3.3. A permissão de uso das imagens contidas no banco de imagens do Ibict não

gera qualquer direito autoral e patrimonial sobre elas.
3.4. É proibido o uso das imagens do acervo para criar conteúdo de caráter difamatório, ilegal, imoral ou obsceno, que possa expor terceiros ao ridículo, violar a moral e os bons costumes ou transmitir informações falsas. Os tipos de ocorrência mencionados podem gerar, ao infrator, penalidades previstas em lei.
3.5. A perda ou o dano de qualquer espécie, ocasionados ao usuário pelo uso devido ou indevido dos recursos do banco de imagens, é de responsabilidade do próprio usuário.
3.6. Também é de responsabilidade do usuário toda e qualquer forma de infração a direitos de terceiros causada pelo uso das imagens contidas no Imago.
3.7. A equipe administrativa do banco de imagens, bem como todo o conjunto de colaboradores do Ibict, não assume responsabilidade sobre a forma de utilização das imagens, incluindo de possíveis resultados danosos.
3.8. A violação de direitos dos autores das imagens disponíveis no acervo do Imago está sujeita às sanções previstas na Lei nº 9.610/98, que protege os direitos autorais no Brasil.

5 Considerações Finais

A utilização de imagens para representação do conhecimento é uma manobra utilizada pela humanidade desde os tempos mais remotos, quando ainda se usava tinta sobre as paredes das cavernas. Desde então, a capacidade de representar mensagens e informações por meio de imagens vem se refinando.

Nos tempos modernos existem inúmeras maneiras de apresentar a informação na forma de imagem: fluxogramas, gráficos, placas, outdoors etc. Sendo assim, a manutenção de um acervo especializado de imagens em diferentes cores e formatos tem grande potencial para facilitar a absorção e disseminação de conhecimento, pois a imagem apresenta maior capacidade de representação quando comparada à escrita.

Denominado Imago, o banco de imagens do Ibict fornece um acervo de eventos, atividades e conteúdos históricos que se relacionam com pesquisa e desenvolvimento de Ciência e Tecnologia. A sua utilização tem como metas a motivação e a disseminação de imagens relacionadas ao Instituto, além de tornar disponíveis ao público os conteúdos imagéticos mencionados.

Este trabalho descreveu como foi o processo de escolha da ferramenta para a formação do banco de imagens do Ibict, denominado Imago. A escolha do software foi guiada por um conjunto de critérios, levantados na fase anterior à prospecção de softwares para bancos de imagens disponíveis na web.

O software escolhido para formar a base de dados do Ibict foi o Piwigo, por ser a única ferramenta a atender todos os critérios levantados e, principalmente, por ser um software livre e viabilizar a manipulação de diferentes grupos de álbuns e usuários.

Por fim, a manutenção e utilização de um banco de imagens também requer um conjunto de diretrizes e políticas para guiar seus usuários no que concerne à reprodução das imagens, bem como seus administradores quanto ao conteúdo disponibilizado ao público por meio do banco.

As políticas de uso dos registros presentes no banco de imagens têm grande importância, visto que são essas normas que regem o uso das imagens pelos seus diversos usuários. Além disso, garantem a utilização correta e íntegra das imagens, sem trazer prejuízo ao Instituto, aos administradores do banco de imagens ou a terceiros.

Para guiar os administradores do banco de imagens do Ibict, foi definido um conjunto de diretrizes e práticas para o gerenciamento do Imago, o qual também compreende informações acerca da licença de acesso e uso das imagens que o compõem.

Referências

DARKTABLE. 2022. Disponível em: https://www.darktable.org/. Acesso em: 17 mar. 2022.

FIGUEIREDO, Márcia Feijão de; SALDANHA, Gustavo Silva. Da linguagem que nos (re) funda ao enigma da imagem nos estudos informacionais: contribuições de teóricos franceses sobre a condição da retórica e imagem na ciência da informação. *In*: XVIII ENCONTRO NACIONAL DE

PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017. **Anais** [...]. Marília, SP: Ancib, 2017. v. 18, p. 1–19. Disponível em: http://hdl.handle.net/20.500.11959/brapci/104491. Acesso em: 17 mar. 2022

GWENVIEW. Disponível em: https://apps.kde.org/gwenview/. Acesso em: 17 mar. 2022.

HAYATI, Sri; SUROSO, Agus; SULIYANTO, Suliyanto; KAUKAB, M. Elfan. Customer satisfaction as a mediation between micro banking image, customer relationship and customer loyalty. **Management Science Letters**, , p. 2561–2570, 2020. DOI 10.5267/j.msl.2020.3.039. Disponível em: http://www.growingscience.com/msl/Vol10/msl_2020_89.pdf. Acesso em: 15 set. 2022.

HOITINK, Yvette. **Wegwijzer opzetten van beeldbanken**. Den Haag: Taskforce Digitale toegankelijkheid Archieven, 2007.

INSTITUTO BRASILEIRO DO MEIO AMBIENTE E DOS RECURSOS NATURAIS RENOVÁVEIS (IBAMA). **Política de uso e privacidade de imagens do banco de imagens do Ibama**. Publicado em 29 de agosto de 2018. Disponível em: http://www.ibama.gov.br/imagens-ibama/368-central-de-conteudos/imagens/banco-de-imagens-do-ibama/1697-politica-de-uso-e-privacidade-de-imagens-do-banco-de-imagens-do-ibama. Acesso em: 24 maio 2022.

IMAGEGLASS. 2018. Disponível em: https://imageglass.org/. Acesso em: 17 mar. 2022.

IMAGO. 2022. Disponível em: http://imago.ibict.br/. Acesso em: 26 maio 2022.

JALBUM. [Homepage]. Disponível em: https://jalbum.net/en/. Acesso em: 17 mar. 2022.

JELLYFIN. [Homepage]. Disponível em: https://jellyfin.org/. Acesso em: 17 mar. 2022.

LIGHTCRAFTS. Disponível em: https://lightzone.softonic.com.br/. Acesso em: 17 mar. 2022.

LYCHEE. [Homepage]. Disponível em: https://lychee.electerious.com/. Acesso em: 17 mar. 2022.

NEXTCLOUD. [Homepage], 2022. Disponível em: https://nextcloud.com/. Acesso em: 17 mar. 2022.

PHOTOPRISM. [Homepage]. 2022. Disponível em: https://photoprism.app/. Acesso em: 17 mar. 2022.

PIWIGO. [Homepage]. 2022. Disponível em: https://piwigo.org/. Acesso em: 17 mar. 2022.

PIWIGO. [Homepage]. 2022. Disponível em: https://piwigo.org/. Acesso em: 24 maio 2022.

RAHI, Samar; GHANI, Mazuri Abd; NGAH, Abdul Hafaz. Factors propelling the adoption of internet banking: the role of e-customer service, website design, brand image and customer satisfaction. **International Journal of Business Information Systems**, v. 33, n. 4, p. 549, 2020. DOI 10.1504/IJBIS.2020.105870. Disponível em: http://www.inderscience.com/link.php?id=105870. Acesso em: 15 set. 2022.

RODRIGUES, Ricardo Crisafulli. Análise e tematização da imagem fotográfica. **Ciência da informação**, v. 36, p. 67-76, 2007.

IMPACTO DA ADEQUAÇÃO À LEI GERAL DE PROTEÇÃO DE DADOS PESSOAIS NA METRIFICAÇÃO DA QUALIDADE DE DADOS

IMPACT OF ADEQUACY TO THE BRAZILIAN GENERAL DATA PROTECTION LAW ON DATA QUALITY METRIFICATION

Leandro Furlam Turi, Giovanni Comarela
Universidade Federal do Espírito Santo, gc@inf.ufes.br

Resumo
A necessidade de avaliar a qualidade de dados vêm ganhando importância, uma vez que a real contribuição (o valor de negócio) dos dados só pode ser estimada no seu contexto de uso. Tais obtenções de vantagens econômicas por meio da mineração de dados sofrem intervenção da Lei Geral de Proteção de Dados Pessoais (LGPD), que define regras sobre o processo de coleta, armazenamento e compartilhamento de informações. Considerando essa relação entre qualidade e legislação corrente, este trabalho avalia e compara o impacto de técnicas de adequação à legislação aos processos de mensuração de qualidade de dados. A partir dos resultados obtidos, notou-se que, em maior ou menor grau, todas as formas de corresponder-se à legislação mostraram-se passíveis a alterações da qualidade da base de dados em relação à base de dados original, demonstrando, assim, que a adequação à legislação deve ser conformada ao fim no qual o projeto se dará.

Palavras-chave: Qualidade de Dados; Lei Geral de Proteção de Dados Pessoais; Ciência de Dados.

Abstract
The demand to assess data quality is gaining importance, since the real contribution (business value) of data can only be estimated in its context of use. Such economic advantages obtained through data mining are subject to the Brazilian General Personal Data Protection Law (LGPD), which defines rules on the process of collecting, storing and sharing information. Considering this relationship between quality and current legislation, this work evaluates and compares the impact of adaptation techniques to legislation on data quality measurement processes. From the results obtained, it was noted that, to a greater or lesser extent, all forms of complying with the legislation proved to be susceptible to changes in the quality of the database in relation to the original database, thus demonstrating that the adequacy to the legislation must be conformed to the purpose in which the project will take place.

Keywords: Data Quality; Brazilian General Personal Data Protection Law; Data Science.

1. Introdução

O acelerado desenvolvimento tecnológico tem modificado a maneira com que se vêm desenvolvendo modelos de negócio no cenário global. Técnicas de processamento e análise de dados têm se mostrado ferramentas bastante usuais na perspectiva do mercado, de modo que os dados vão se constituindo no cerne da tomada de decisão e do planejamento estratégico de empresas (ARDAGNA et al, 2008).

Estratégias de mineração de dados estão surgindo como uma nova solução para os problemas encontrados ao se processarem grandes quantidades de dados. O problema é que nem todos os dados são padronizados: “uma das dificuldades fundamentais é que as informações extraídas podem ser tendenciosas, ruidosas, desatualizadas, incorretas, enganosas e, portanto, não confiáveis” (BERTI-EQUILLE, L.; BORGE-HOLTHOEFER, 2015).

As obtenções de vantagens econômicas por meio da mineração de dados também sofrem intervenção da Lei Geral de Proteção de Dados Pessoais (LGPD) (PINHEIRO, 2020). Essa Lei define regras sobre o processo de coleta, armazenamento e compartilhamento de informações, de forma a fazer valer observância à boa-fé e aos demais princípios elencados no Art. 6º, com especial destaque para os critérios de finalidade (tratamento com propósitos legítimos), prevenção de potenciais danos ao titular dos dados e responsabilização de danos patrimoniais ou morais causados decorrente da atividade de tratamento de dados pessoais, de modo a coibir abusos e

favorecer um ambiente mais seguro aos usuários.

Considerando a relação entre qualidade e legislação corrente, este trabalho avalia e compara o impacto de técnicas de anonimização em acordo com a Lei Geral de Proteção de Dados Pessoais aos processos de mensuração de qualidade de dados. Para isto, serão levantados conjuntos de dados pertencentes a diferentes contextos, aos quais serão aplicadas diferentes técnicas de anonimização acordadas à LGPD. Com cada conjunto anonimizado, é calculada a qualidade dos dados com métricas definidas pela ISO/IEC 25012:2008, cujos cômputos serão comparados e avaliados em relação aos resultados obtidos com a aplicação à base original. Com este processo, pretende-se responder a seguinte pergunta de pesquisa: Como a LGPD influencia na mensuração da qualidade de uma base de dados? Isto reflete-se em projetos de ciência de dados?

2. Objetivos

Analisar a viabilidade e a adaptabilidade de processos de mensuração de qualidade de dados através de diferentes formas de adequação à LGPD: (1) categorizando as abordagens e os requisitos gerais atuais para cada métrica a ser analisada, demonstrando estruturas e desafios; (2) implementando e automatizando tais técnicas em conjunto com algoritmos de suporte para o processo de qualidade de dados; (3) comparando e avaliando os resultados obtidos com cada técnica de anonimização em relação à base original.

3. Procedimentos Metodológicos

A metrificação da qualidade de dados constitui-se de um esquema cíclico, iniciando no mapeamento da documentação e do comportamento dos dados, através da observação de trechos das bases. Em seguida, são definidas as variáveis de teste. Após, ocorre a obtenção e avaliação dos resultados obtidos, recorrendo, e se necessário retificando, conclusões obtidas nos passos anteriores (MERINO, 2016). Todos os códigos utilizados, bem como a base de dados apresentada, estarão disponíveis de forma pública na publicação do trabalho.

Embora a literatura apresente muitas dimensões de qualidade de dados, trazer à tona padrões internacionais como ISO/IEC 25012:2008 e ISO/IEC 25024:2015 pode ser muito conveniente, e eles podem ser utilizados como guias de referência. A ISO/IEC 25012:2008 contém um modelo de qualidade de dados com um conjunto de características que os dados de qualquer sistema de informação devem cumprir para atingir níveis adequados de qualidade de dados externos. A ISO/IEC 25024:2015 fornece medidas gerais para quantificar a qualidade externa e a interna dos dados em conformidade com características da ISO/IEC 25012:2008. Nesse sentido, seis métricas inerentes a sistemas foram selecionadas: *Completude*, *Acurácia*, *Consistência*, *Credibilidade*, *Atualidade* e *Unicidade*. Uma descrição de cada métrica é apresentada a seguir, enquanto exemplos de aplicação poderão ser encontrados na discussão de resultados.

Completude caracteriza a taxa de preenchimento dos atributos. *Acurácia* visa a detectar se a informação registrada reflete o evento ou o objeto descrito, isto é, verificar se o dado cadastrado está em concordância com o evento observado. Tem três aspectos principais: (1) Acurácia sintática, definida como a proximidade dos valores dos dados a um conjunto de valores definidos em um domínio considerado sintaticamente correto; (2) Acurácia de alcance, que define os intervalos nos quais os valores dos dados devem ser definidos em um domínio considerado semanticamente correto; (3) Acurácia Semântica, definida como a proximidade dos valores dos dados a um conjunto de valores definidos em um domínio considerado semanticamente correto. *Consistência* avalia se dois ou mais atributos estão livres de contradição e são coerentes com outros dados de contextos coerentes. *Credibilidade* avalia se os atributos são considerados verdadeiros e críveis pelos usuários. *Atualidade* avalia se os dados representam o período factível e coerente. *Unicidade* mensura o grau de duplicidade nos dados.

O cômputo dos resultados numéricos ocorre de modo cascata, gerando-se a cada nível um domínio de aplicação. Por exemplo, em um fluxo completo contendo todas as

métricas em uso, os registros submetidos ao teste de acurácia devem ser não nulos; os registros submetidos ao teste de credibilidade devem estar acurados; os registros submetidos aos testes de consistência deverão ser credíveis; ao teste de atualidade devem ser consistentes; e os registros submetidos aos testes de unicidade devem ser atuais. Destaca-se que, quando não for possível metrificar para uma variável específica, a métrica antecedente à que seria utilizada, a nível de registro, deve ser trazida para definir o domínio de aplicação.

Acerca das bases de dados analisadas, procurou-se levantar exemplos de conjuntos de dados nacionais e internacionais em que determinadas informações pessoais foram publicadas como parte do registro público. A inspiração partiu de um tópico[1] da *Open Knowledge Foundation* e pelos tuítes de *@jwyg*, onde a discussão acerca de privacidade decorreu. A saber, as bases utilizadas foram: *athleteEvents*, *Canada*, *eleicoes*, *EuropeanSoccer*, *Poland* e *Sinasc*. A natureza e o tratamento realizado de cada base está disponibilizado no repositório aberto vinculado à pesquisa[2].

Destaca-se que não necessariamente as informações obtidas são objetos da LGPD. Entretanto, como claramente caracterizam-se como informações pessoais (mesmo que referindo-se a pessoas públicas, excluídas da legislação), servem como exemplo para a proposta deste trabalho.

Retornando à adequação à legislação vigente, a LGPD inclui na lista de dados sensíveis aqueles que dizem respeito à "origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico, quando vinculado a uma pessoa natural". Uma vez definidos os objetos da anonimização, basta definir os mecanismos. O inciso XI do Art. 5º define anonimização como sendo "utilização de meios técnicos razoáveis disponíveis no momento do tratamento, por meio dos quais um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo". Buscando eliminar tais elementos identificadores, o trabalho foi feito sobre quatro grandes técnicas conhecidas em computação: *Supressão (SUP)*, *Generalização (GEN)*, *Randomização (RAND)* e *Pseudoanonimização (PSAN)*.

Supressão consiste na ação de suprimir (ou anular) determinada variável referindo-se a um registro. Por exemplo, excluir os dígitos de um número de telefone ou todos os nomes de uma base de dados. *Generalização* consiste em substituir os registros específicos por categorias mais amplas e genéricas. Por exemplo, idades exatas são convertidas em faixas etárias e um CEP é trocado apenas pela cidade ou região do país. *Randomização* é o processo de tornar o registro aleatório. Neste trabalho, os registros de cada variável submetida a esse processo são permutados aleatoriamente. *Pseudoanonimização* consiste no mecanismo do encobrimento da informação, substituindo-se um atributo por outro. Neste trabalho, o hash SHA224[3] foi utilizado.

4. Resultados

4.1 Completude

Nessa dimensão são detectados valores faltantes mediante a busca por tipos ou valores representando a informação nula na base de dados. Destaca-se que para esta métrica faz-se necessário observar a dependência entre as variáveis definidas nos dicionários de dados adotados, como por exemplo a existência de uma anomalia em contrapartida ao código representativo desta (*Sinasc*). A Fig. 1 apresenta a completude de cada base de dados analisada para cada aplicação de técnica de anonimização, com o respectivo percentual numérico. Destaca-se a grande queda de completude dos registros para a técnica de supressão, em todas as bases analisadas, seguida pela queda para a técnica de generalização.

A queda para a supressão decorreu do fato de que, durante a ação de anular determinadas variáveis, essas passam a ser incompletas, causando as quedas observadas. Já para a generalização, a queda decorreu da aplicação da técnica de forma total para as variáveis.

[1] *https://lists-archive.okfn.org/pipermail/mydata-open-data*

[2] *https://github.com/leandrofturi/qualityLGPD*

[3] *https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf*

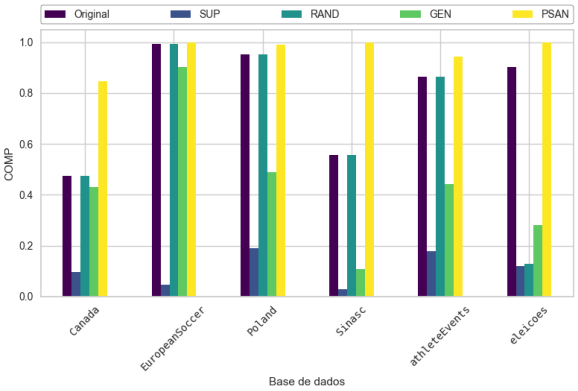


Figura 1 – Resultados para métrica de completude

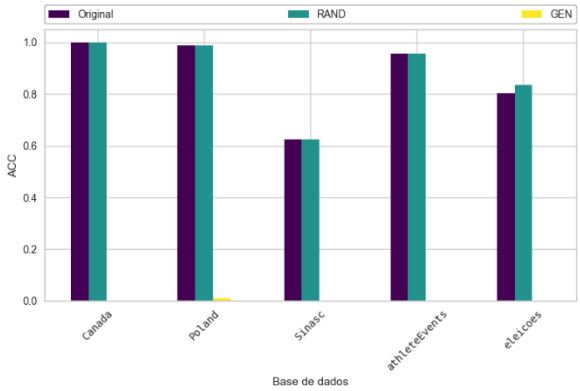


Figura 2 – Resultados para métrica de acurácia

4.2 Acurácia

Verificou-se se os dados apresentam os padrões descritos no dicionário de dados adotado como referência, tanto de quantidade de caracteres, quanto de valores esperados, em três aspectos principais: (1) sintático, definido como a proximidade dos valores dos dados a um conjunto de valores definidos em um domínio considerado sintaticamente correto. Isso abrange todos os domínios finitos definidos pelo dicionário de dados, bem como formato dos valores (também conhecido como conformidade); (2) alcance, definido pelos intervalos aos quais os valores dos dados devem ser definidos, abrangendo principalmente datas; e (3) semântico, definido como a proximidade dos valores dos dados a um conjunto de valores definidos em um domínio considerado semanticamente correto, definido pelos dicionários de dados, abrangendo principalmente pesos, idades e medidas dentro do contexto no qual a base de dados se encontra. O resultado percentual por variável está descrito na Fig. 2.

Destaca-se a grande queda de acuracidade dos registros para a técnica de generalização, uma vez que várias verificações, que antes eram passíveis de serem aplicadas na base de dados original, pela aplicação não são mais necessárias. Isso inclui testes envolvendo conformidades ao dicionário de dados, faixas de valores para datas e verificações de idades e acuracidade nos códigos de cidades e municípios.

4.3 Credibilidade

Verificou-se o grau em que os dados têm atributos que são considerados verdadeiros e críveis, incluindo o conceito de autenticidade (a veracidade das origens, atribuições, compromissos) ISO/IEC 25012:2008. Nesse contexto, testes envolvendo a origem da informação (municípios de nascimento para a base de dados *Sinasc* e unidade eleitoral para a base de dados *eleicoes*) foram realizados, objetivando verificar se de fato a informação disponibilizada refere-se ao recorte realizado (estado do Espírito Santo). Poucos foram os problemas identificados, conforme apresentado na Fig. 3.

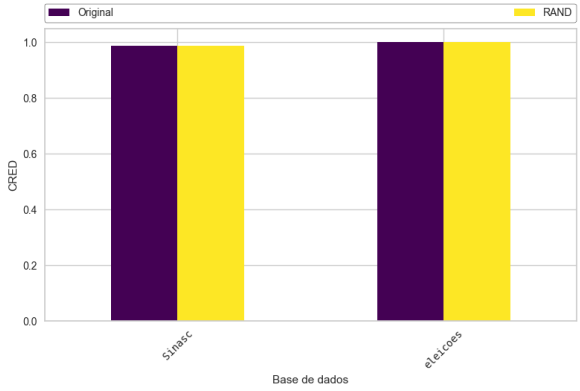


Figura 3 – Resultados para métrica de credibilidade

4.4 Consistência

Os resultados descritos na Fig. 4 são de testes aplicados a um mesmo registro, ou seja, mesma linha do conjunto de dados. Estes testes detectam principalmente problemas na entrada de dados envolvendo condições específicas de inconsistências e dependentes do contexto. Por exemplo, para a base de dados *Sinasc*, um teste envolvendo a data de nascimento do recém-

nascido e a data de nascimento da mãe varre os registros identificando se a data do nascimento do recém-nascido é maior que a data de nascimento da mãe. Todos os demais testes aplicados estão descritos no repositório aberto desta pesquisa.

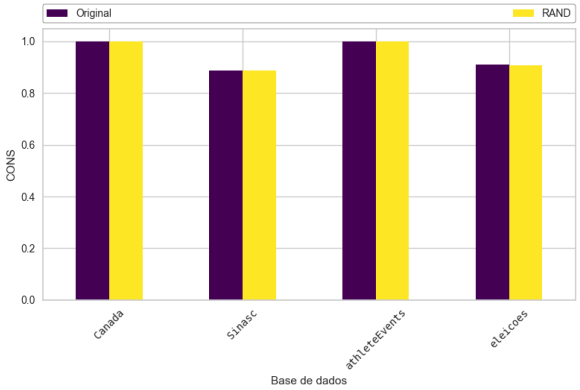


Figura 4 – Resultados para métrica de consistência

4.5 Atualidade

Nessa métrica foi verificado o grau em que os dados têm atributos que estão na temporalidade adequada para o uso. Poucos testes foram identificados: em *EuropeanSoccer*, verificar se cada registro possui informações contínuas (sem lacunas entre anos); e em *Sinasc*, se a informação do nascimento era submetida ao sistema em no máximo um ano. Os resultados são apresentados na Fig. 5.

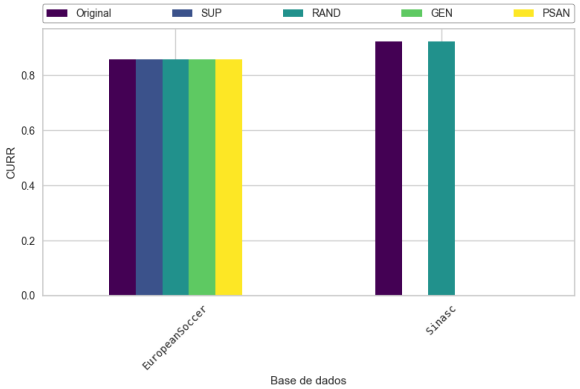


Figura 5 – Resultados para métrica de atualidade

4.6 Unicidade

Por fim, na métrica de unicidade foi analisado o grau de duplicidade dos registros, por meio da identificação do registro por um identificador único e pela verificação das constantes que definem os registros, como por exemplo nome, data e local de nascimento. Os resultados percentuais são apresentados pela Fig. 6.

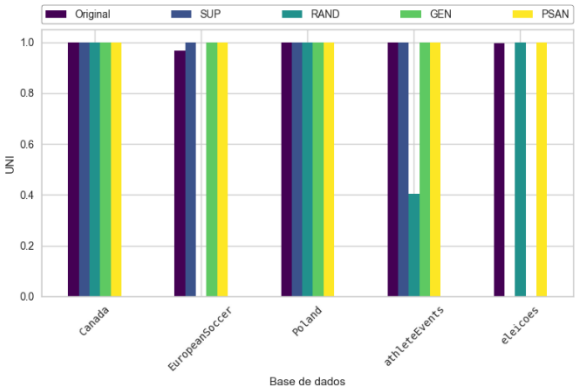


Figura 6 – Resultados para métrica de unicidade

De fato, a técnica de randomização é a mais favorável a apresentar erros quando existe mais de uma linha relativa a um mesmo registro, uma vez que essa técnica irá embaralhar a ordem das informações. Por outro lado, as técnicas de supressão, generalização e pseudoanonimização são favoráveis a manter a quantidade de falhas, uma vez que as variáveis se tornam mais amplas, e menos susceptíveis a erros. Tome por exemplo o campo referindo-se ao código de seis dígitos do município de nascimento do recém-nascido vivo de acordo com a tabela de códigos e municípios do IBGE[4], na base de dados *Sinasc*. Generalizando-se essa informação, ou seja, mantendo-se apenas os dois primeiros dígitos referentes apenas ao estado de nascimento, a possibilidade de haver erros nessa informação é reduzida. Isso vale para os demais testes.

5. Considerações Finais

A avaliação realizada é especialmente oportuna, tendo em vista o cenário nacional e o atual empenho em fomentar o debate em torno da qualidade das informações e o respeito à privacidade do brasileiro. Nesse contexto, este trabalho explorou métricas de qualidade de dados presentes em normas técnicas, objetivando avaliar a qualidade dos dados de conjuntos de diferentes contextos e o impacto que a adequação à LGPD ocasiona nestes projetos. Para isso, foram metrificadas as qualidades das bases de dados originais (*athleteEvents*, *Canada*,

[4] *https://www.ibge.gov.br/explica/codigos-dos-municipios.php*

eleicoes, *EuropeanSoccer*, *Poland* e *Sinasc)*, e comparadas com o resultado obtido após conformação com a legislação.

Analisados os resultados obtidos pela métrica de completude, observou-se uma grande queda para a técnica de supressão, em todas as bases analisadas, seguida pela queda para a técnica de generalização. A queda para a supressão decorreu do fato de que, durante a ação de anular determinadas variáveis, estas passam a ser incompletas, causando os resultados observados. Já para a generalização, a queda decorreu da aplicação da técnica de forma completa para algumas variáveis.

Sobre os resultados dos testes de acurácia, destacou-se a grande queda de acuracidade dos registros para a técnica de generalização, uma vez que várias verificações que antes eram passíveis de serem aplicadas na base de dados original, pela generalização não são mais necessárias. Isso inclui testes envolvendo conformidades ao dicionário de dados, faixas de valores para datas e verificações de idades e acuracidade nos códigos de cidades.

Com relação aos resultados de credibilidade, consistência, atualidade e unicidade, observou-se pouca variação da metrificação, quando a técnica aplicada conseguiu manter a semântica da informação, mesmo que mascarada.

Baseado nestes resultados, nota-se uma técnica que se sobressai às demais, a saber, a pseudoanonimização. Uma vez que esta altera apenas a sintaxe dos registros (os valores visíveis, da forma como são escritos), não se deve esperar que haja diferenciações na aplicação desses algoritmos, desde que, claro, as informações originais não sejam de maioria numérica (perdendo-se, com a aplicação da técnica, a semântica intrínseca da informação). Tal fato foi tanto levantado na metrificação da qualidade e pode ser elencado em projetos posteriores de ciência de dados.

Visando observar a fundo tal comportamento evidenciado pela técnica de pseudoanonimização, realizar análises via algoritmos de aprendizado de máquina (de natureza geométrica como um *KMeans*[5]) em que a semântica intrínseca aos registros não seja requisito para o bom funcionamento podem ser redigidas e avaliadas.

Referências

ARDAGNA, D. et al. Context-aware data quality assessment for big data**. Future Generation Computer Systems**, v. 89, p. 548–562, 2018. ISSN 0167-739X.

BERTI-EQUILLE, L.; BORGE-HOLTHOEFER, J. Veracity of data: From truth discovery computation algorithms to models of misinformation dynamics. **Synthesis Lectures on Data Management**, Morgan & Claypool Publishers, v. 7, n. 3, p. 1–155, 2015.

ISO/IEC JTC 1/SC 7 Software and systems engineering technical committee. **ISO/IEC 25012:2008 Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model**. 1. ed. Vernier, Geneva, Switzerland, 2012.

ISO/IEC JTC 1/SC 7 Software and systems engineering technical committee. **ISO/IEC 25024:2015 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality**. 1. ed. Vernier, Geneva, Switzerland, 2015.

PINHEIRO, P. P. **Proteção de Dados Pessoais: Comentários à Lei n. 13.709/2018-LGPD**. 2. ed. São Paulo, São Paulo, Brasil: Saraiva, 2020. ISBN 9788553617487.

MERINO, J. et al. A data quality in use model for big data. **Future Generation Computer Systems**, v. 63, p. 123–130, 2016.

[5] *http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf*

LIBERDADE DE EXPRESSÃO: A NARRATIVA NO TWITTER EM UM CONTEXTO DE ANÁLISE DE REDES SOCIAIS

FREEDOM OF EXPRESSION: THE NARRATIVE ON TWITTER IN A CONTEXT OF SOCIAL NETWORK ANALYSIS

Stella Schwanz Dias de Assis[1], Meri Nadia Marques Gerlin[2]
(1) Universidade Federal do Espírito Santo, stella.assis@edu.ufes.br
(2) Universidade Federal do Espírito Santo, meri.gerlin@ufes.br

Resumo
O ciberespaço proporciona a circulação de informações com uma intensidade como nunca antes vista. Em um contexto do ciberespaço, é possível de se observar diversas mudanças na dinâmica das relações de influência nas plataformas de mídias sociais, onde são atribuídos valores a elementos que antes não possuíam significativa importância. Com o objetivo analisar as relações de influência e compreender o papel das autoridades em uma rede social, a presente pesquisa desenvolve uma análise qualitativa com auxílio de tratamentos quantitativos de dados abertos do Twitter. A pesquisa combina procedimentos de Análise de Redes Sociais com Estudos de Discurso mediados por computador, a partir dos dados coletados no Twitter com o auxílio das ferramentas Netlytic para coleta de dados e Gephi para elaboração dos grafos. Nos resultados observa-se que com a análise baseada na representação da informação através de grafos, em um contexto da Web Semântica, foi possível observar a influência que as autoridades digitais podem exercer diante de um debate nas plataformas de mídias sociais.

Palavras-chave: Análise de Redes Sociais; Liberdade de Expressão; Twitter.

Abstract
Cyberspace provides a circulation of information with an intensity like never seen before, in a context where it is possible to observe several changes in the dynamics of influence relations on social media platforms, where values are attributed to elements that previously did not have significant importance. With the aim of analyzing influence relationships and understanding the role of authorities in a social network, this research develops a qualitative analysis, with the aid of quantitative treatments of open data from Twitter. The research combines Social Network Analysis procedures with computer-mediated Discourse Studies, based on data collected on Twitter with the help of Netlytic tools for data collection and Gephi for drawing up graphs. In the results, it is observed that with the analysis based on the representation of information through graphs, in a context of the Semantic Web, it was possible to observe the influence that digital authorities can exert in the face of a debate on social media platforms.

Keywords: Social Network Analysis; Freedom of expression; Twitter.

1. Introdução

O ciberespaço proporciona uma circulação de informações com uma intensidade como nunca antes vista. Destaca-se a importância de melhor compreender a dinâmica das redes que se constroem nas mídias sociais, justamente pelo impacto que essas construções podem causar na sociedade. Por se tratar de um contexto de plataformas de mídias sociais, que possuem um amplo fluxo de informação, uma grande quantidade de usuários e pouco monitoramento frente ao conteúdo compartilhado, fazem-se necessários diferentes recursos e estratégias, pautados na metodologia de análise de redes e análise de conteúdo no contexto da Web Semântica, a fim de compreender as relações de influência de uma rede constituída a partir da recuperação da informação com base na coleta de dados organizados na base do Twitter.

Em um contexto do ciberespaço, é possível observar diversas mudanças na dinâmica das relações de influência, onde são atribuídos valores a elementos que antes não possuíam significativa importância.

Analisando a organização dos indivíduos nas mídias sociais pode-se enxergar essa formação de grupos, construídos a partir de aglomerados de ideologias, onde, cada vez mais, as ideias similares podem se aproximar e fortalecer. Com as tecnologias digitais, os indivíduos passaram a possuir um acesso imensurável à informação, no entanto tendem a buscar aquelas que fazem parte de um escopo que já acreditam e estão familiarizados. Tendência que é reforçada por algoritmos que levam informações de acordo com os interesses de cada usuário.

Ao entrar em contato com uma grande quantidade de perfis, que podem ser indivíduos ou até robôs, que compartilham de seus ideais, algumas visões acabam se intensificando pelo sentimento de pertencimento.

O método de análise de redes sociais (ARS), vem sendo amplamente utilizado em pesquisas nas mídias sociais digitais, buscando compreender discussões específicas que acontecem nestes meios de comunicação. As mídias sociais estão cada vez mais presentes no cotidiano das pessoas, deixando de ser algo secundário. Vinhas, Sainz e Recuero (2020, p.41) apontam que existe uma complexidade nas interações sociais na internet, por conta de "um intrincamento permanente entre o que ocorre no real e os aspectos específicos atinentes às relações constituídas no digital."

A influência das mídias sociais transita por diferentes aspectos da sociedade, não influenciando apenas no desejo de compra de um indivíduo, mas em seu comportamento, suas decisões, até mesmo na escolha do voto em período eleitoral. Paulino e Ventura (2021, p. 68) evidenciam que "vivemos um momento de descoberta do poder de engajamento das mídias sociais", onde "as pessoas estão mobilizadas e engajadas para se unirem em torno de pautas em comum e mostrar seu descontentamento". Por essas questões, métodos que contribuem para a compreensão das relações de influência e poder nesses meios se fazem tão importantes.

A infinidade de dados disponíveis na Web tornou o procedimento de classificação da informação com base na qualidade e relevância extremamente importante para o processo de organização e recuperação de dados. Em uma mediação através do raciocínio humano, essa classificação já é compreendida, mas os princípios da Web Semântica se fazem essenciais ao analisar a forma que a interpretação dessas informações é realizada pelas máquinas, o que determina a dinâmica da sua disponibilização para os usuários (NUNES, MACULAN, ALMEIDA, 2020).

Recuero, Bastos e Zago (2015, p.14), destacam que a Análise de Redes contribui na classificação de atores em uma população a partir de sua localização na rede, simplificando esse processo através de métricas aplicadas nos dados coletados. Essas métricas e indicadores da Análise de Redes Sociais (ARS) são essenciais na compreensão dos atores mais relevantes, representados pelos nós com maior capacidade de dinamizar e expandir as relações dentro do escopo delimitado na pesquisa. (KALINE e ROCHA, 2020; RECUERO e GRUZD, 2019).

É interessante destacar que cada mídia social possui um padrão de comportamento de usuários diferente. O Twitter foi selecionado nesta pesquisa por possuir um fluxo mais dinâmico, rápido. Recuero e Soares (2021, p.6) compreendem "a disputa discursiva no Twitter como um embate pela legitimidade". Os autores destacam que essa legitimidade é obtida através de métricas de reprodução. A importância não estaria no sujeito ou nos enunciados, mas no número de *retweets,* comentários, curtidas, menções a um *tweet* original, ações que, nestes ambientes, demonstram apoio a determinada formação discursiva. Araújo, Moraes e Pisa (2020, p.19) destacam que, mesmo contando com publicações com textos curtos, onde existe um limite baixo de caracteres, "a análise exploratória de grafos em um conjunto grande de mensagens é bastante útil para compreensão de uma narrativa ampla, auxiliando na identificação dos assuntos discutidos de forma simplificada".

A Análise de Redes nas Mídias Sociais não permite apenas o mapeamento e investigação do significado de determinada postagem, mas possibilita a identificação de diversas informações, assim como "verificar tendências, mapear polaridades, verificar sentimentos, níveis de toxidade e muitos outros fenômenos", através da análise das redes formadas pelos usuários e por suas mensagens (EMPINOTTI et al., 2021).

2. Objetivos

O presente estudo traz um foco sobre uma discussão centrada na liberdade de expressão no Twitter, com o objetivo de analisar as relações de influência compreendendo o papel das autoridades em uma rede social. Para isso, a presente pesquisa desenvolve uma análise qualitativa, com auxílio de tratamentos quantitativos de dados abertos do Twitter, buscando compreender como são estruturadas as redes sociais que compõem a discussão relacionada à liberdade de expressão,

delimitando as principais clusters e analisando as relações entre as principais temáticas que são relacionadas à Liberdade de Expressão dos tweets coletados.

3. Procedimentos Metodológicos

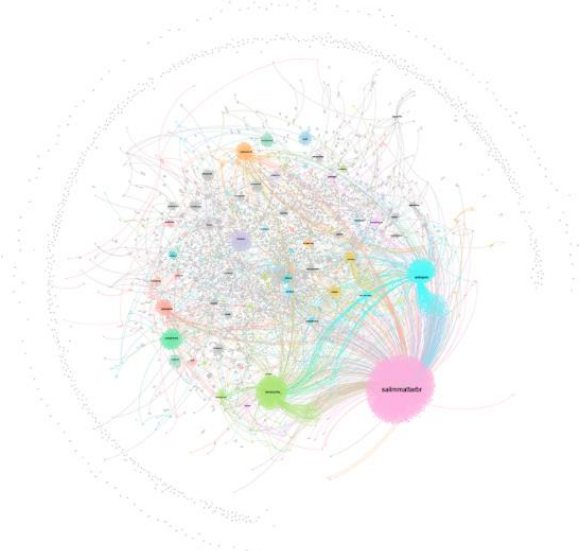
A presente pesquisa desenvolve uma análise qualitativa, com auxílio de tratamentos quantitativos de dados. Combina procedimentos de Análise de Redes Sociais com Estudos de Discurso mediados por computador, a partir da coleta na base de dados abertos do Twitter, identificando inicialmente as conexões entre os perfis, para depois levantar os tópicos mais relevantes identificados nos textos das postagens e as conexões entre esses termos.

Para investigar a rede formada pela discussão da temática central, foi utilizado o software Netlytic[1] para recuperação dos dados no Twitter, através da busca pelo termo "liberdade de expressão". Essa plataforma oferece um relatório com até dez mil postagens no período de até 7 dias antes da busca, gerando uma base no formato CSV com todos os *tweets*, o gráfico visual da rede, nuvem de palavras, além de um relatório geral, com dados como os usuários que mais postaram sobre o tema. A coleta de dados foi realizada no dia 10 de dezembro de 2021, coletando as postagens publicadas entre o dia 02 e dia 10 de dezembro. Foram coletados 10.000 *tweets* e 8.440 postagens únicas.

Como o *Netlytic* possui recursos limitados, os dados foram exportados para o software *Gephi*, onde é possível realizar uma série de filtragens e reorganização do gráfico. Com os dados exportados foram identificados 9.930 nós e 8.986 arestas. O primeiro passo foi a definição das métricas de nó, onde foi utilizada como parâmetro de tamanho, a Centralidade de Autovetor, que leva em consideração as conexões diretas e indiretas, focando na influência dos nós na rede (RECUERO; BASTOS; ZAGO, 2015, p.72). Buscando um maior destaque das comunidades, foi executado nas métricas de rede o Grau de Modularidade, que calcula os nós que estão mais densamente interligados para agrupamento em módulos (RECUERO; BASTOS; ZAGO, 2015, p.84). Essa métrica foi aplicada como atributo de distribuição de cor na representação dos nós e arestas.

O layout MultiGravity ForceAtlas 2 foi aplicado na definição da disposição do gráfico. O algoritmo ForceAtlas foi selecionado por computar o grau de cada nó na força de repulsão da distribuição, possibilitando uma melhor visualização das comunidades da rede (RECUERO; BASTOS; ZAGO, 2015, p.103), resultando na representação retratada na Figura 1.

Figura 1- Visualização da rede de pessoas conectadas que utilizaram o termo "Liberdade de Expressão" no período pesquisado, 2021



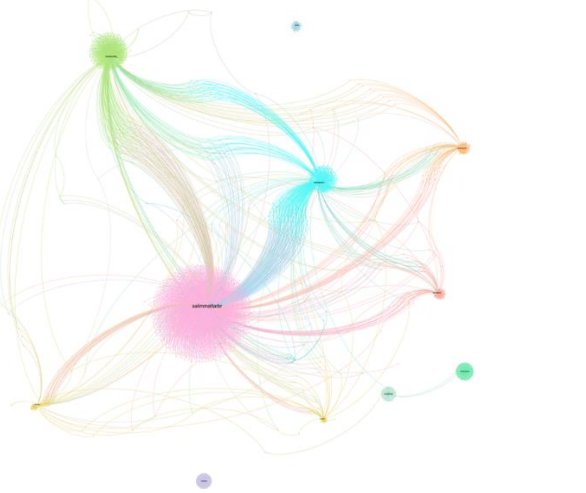
Fonte: Grafo produzido pelas autoras no software Gephi.

4. Resultados

A fim de melhor representar os nós mais fortemente conectados da rede, destacando as principais comunidades formadas, foi aplicado um filtro de rede de vizinhos, utilizando como parâmetro o intervalo de centralidade de autovetor maior que 0.009. Em seguida, afim de proporcionar uma visualização mais nítida dos principais clusters, formados pelas conexões com os nós de maior relevância na rede, foi alterado o intervalo de centralidade de autovetor do filtro já aplicado, considerando somente aqueles que são superiores a 0.03. Na Figura 3 é possível enxergar a formação de 11 comunidades, que se concentram em torno de atores específicos, destacando as interconexões entre clusters, onde algumas possuem diversos nós em comum e outras possuem pouca, ou até mesmo, nenhuma conexão com as demais.

[1] Software para análise de redes sociais, disponível em: https://netlytic.org/

Figura 3 - Visualização dos nós mais relevantes da rede e suas conexões



Fonte: Grafo produzido pelos autores no software Gephi.

Com base nessa identificação dos principais atores da rede, foram selecionados esses nós centrais das 11 comunidades, exportando os dados dos mesmo para uma planilha CSV, possibilitando uma análise mais detalhada desses atores. Na tabela 1 é possível observar a descrição do principal tweet publicado por cada usuário, além de dados como o número de seguidores, a bio do perfil, destacando também os perfis que são verificados pela plataforma.

Quadro 1 - Perfis com maior centralidade de autovetor

Usuário	Seguidores	Bio	Descrição	Verificado	Eigencentrality
salimmattarbr	183626	Ideias liberais para um Brasil melhor!	Presidente do TSE, ministro Barroso, disse que é imperativo a regulação das redes sociais. Não existe democracia, sem plena liberdade. Regular a liberdade de expressão coloca a democracia em risco.	SIM	1
nonaconta_	8080	Aqui a paciência é curta, e sobra personalidade. Salmo 91 No DM Agindo Deus quem impedirá? #BolsonaroAte2026 e continua LALALÚ	Todos nós deveríamos fazer isso!!! Denunciar internacionalmente todos os ministros que cometem arbitrariedades e abusos contra a nossa liberdade e o direito de livre expressão!!! https://t.co/V29ezSatDt	NÃO	0,312904
andreporci	114893	Secretário Nacional de Incentivo e Fomento à Cultura - Lei Rouanet / Capitão da PM-BA	Li a denúncia apresentada na Comissão Interamericana de Direitos Humanos, sobre uma inexistente violação da liberdade de expressão artística. É uma piada completa, não consigo entender como a Comissão marcou uma audiência pública sobre isso. Enfim, prestarei os esclarecimentos.	SIM	0,253393
carapanarana	90020	Cronista do Armagedom / Palpiteiro no podcast @viracasacas https://t.co/IOGL3e2VeN	Muito engraçado como toda essa turma contra o "cancelamento" e a favor da "liberdade de expressão" tá caladinha com a possibilidade de um projeto de lei que cria uma milícia particular do bolsonarismo pra perseguir e encarcerar toda e qualquer oposição.	NÃO	0,086368
342artes	8888	Contra a censura e a difamação. #342ARTES	. @caetanoveloso, @danielamercury e Wagner Moura participarão da audiência junto com integrantes do @movimentomobile falando sobre as violações de liberdade de expressão no país, o desmonte institucional do setor cultural e a perseguição de artistas. https://t.co/laPeHTG1zE https://t.co/73YuX3xhWd	NÃO	0,084772
capitaoderrite	176500	Deputado Federal por SP 2018, Oficial de ROTA (2010/2013), Oficial do Corpo de Bombeiros (2015/2018).	O tal "ódio do bem". Não se pode questionar, muito menos ir contra a imposição deles. Os que mais se dizem defensores da liberdade de expressão são os primeiros a tentar calar vozes contrárias. CRIMINOSOS!	SIM	0,078617
generalgirao	187919	General da Reserva do Exército Brasileiro. Deputado Federal pelo Rio Grande do Norte. BRASIL acima de tudo!	A ridícula intolerância religiosa sofrida pela primeira-dama, @MiBolsonaro , só comprova o que já sabíamos: No Brasil, quando se trata de cristãos, a intolerância é relativizada como liberdade de expressão. Sim, Somos pacíficos, mas também TEMPLÁRIOS. Deus Vult. \ † #RespeiteAFe	SIM	0,06693
liacrespo	55643	Robovó a favor do Brasil e contra os bocós. "Nunca se renda, exceto às convicções de honra e ao bom senso." Winston Churchill	#OsPingosNosIs "O ministro Barroso está convencido de que pode deixar a realidade pra fora. A liberdade de expressão não interessa mais a essa corte a que pertence esse personagem." @GFiuza_Oficial	NÃO	0,041598
philliphonorato	13654	Meu Hobby é fazer montagem mal feita e ninguém achar graça. Temos de nos tornar a mudança que queremos ver no mundo.	Viva a liberdade de expressão!!! Só não pode criticar o Bolsonaro https://t.co/yE1YhiFjCr	NÃO	0,034106
lindasod	18458	Jornalista, Produtora, Tuiteira, Interessada em Política e Patriota. Amo meu Brasil em toda a sua Potência. Com pessoas mal educadas, não tem conversa é block.	Convidado por Joe Biden, Bolsonaro participa, hoje, da reunião de líderes mundiais: A "Cúpula pró-Democracia". No discurso, Bolsonaro irá frisar seu posicionamento pela internet livre, fazer críticas a imprensa tradicional e o compromisso de salvaguardar a liberdade de expressão. https://t.co/7Iw3GYu4dM	NÃO	0,031877
pco29	28871	O PCO é um partido revolucionário e comunista, que baseia sua luta na organização e no desenvolvimento da classe trabalhadora.	A liberdade de expressão é um direito fundamental, ela é na realidade a liberdade do pensamento. Quem defende limites para esse direito esta defendendo a ditadura, ou seja, a repressão do Estado burguês contra os trabalhadores e suas organizações.	NÃO	0,03138

Fonte: Produzido pelos autores com base nos dados da pesquisa.

No Quadro 1 podemos observar uma comunidade que se destaca tanto pela quantidade de atores envolvidos, quanto por sua centralidade na rede, com interconexões com 6 dos demais clusters. Representado na cor rosa, observamos o agrupamento que tem como ator central o perfil de Salim Mattar, um dos maiores defensores da causa liberal no Brasil, em sua principal postagem da amostra articula a liberdade de expressão com os tópicos democracia e regulação das redes sociais, em crítica ao Ministro Luís Roberto Barroso.

Essa comunidade, juntamente com as visualizadas em azul e verde claro, representam os clusters mais intensamente interconectados na rede. No verde observamos como ator central o perfil "nonaconta_", centrado em discussões políticas e apoio ao Bolsonaro, sua principal publicação defende a denúncia do Ministro Alexandre de Moraes. Já em azul, observa-se como ator central o Secretário Nacional de Incentivo e Fomento à Cultura, André Porciuncula, que no tweet em destaque critica a denúncia de violação à liberdade de expressão artística apresentada na Comissão Interamericana de Direitos Humanos.

Também é identificada uma intensa interconexão com as clusters que tem como ator principal, o Deputado Federal de São Paulo Capitão Derrite (laranja), o Deputado Federal pelo Rio Grande do Norte General Girão Monteiro (salmão), o perfil de Lia Crespo, ativista pró-governo (amarelo, localizado à esquerda) e o perfil @lindasod, que se autointitula como jornalista interessada em política e patriota (amarelo, localizado à direita). Com base nos dados coletados, e investigando os perfis mencionados, foi observado que os mesmos possuem em comum um posicionamento pró-governo, centralizando a discussão na defesa de uma liberdade de expressão plena e sem limitações, assim como os três maiores clusters citados anteriormente. Neste contexto de perfis com posicionamentos políticos similares, apenas dois dos principais tweets apresentam uma abordagem divergente das demais, publicados pelos perfis do Capitão Derrite e do General Girão, associando a liberdade de expressão como premissa para silenciamento e relativização da intolerância de posicionamentos contrários.

Dentre as comunidades menos conectadas, observa-se, na parte inferior à direita do grafo, a centrado no perfil "philliphonorato", que possui interconexões com dois clusters do grupo citado anteriormente, além de ser o único a possuir nós em comum com o perfil "carapanarana". Em ambos perfis foi observado um posicionamento contrário ao governo, as postagens principais dos usuários possuem até mesmo uma abordagem similar, onde

criticam apoiadores do Bolsonaro por posicionamentos incoerentes ligados à liberdade de expressão.

Os dois últimos clusters, são encontrados isolados dos demais, sem interconexões. Em azul podemos observar as conexões com o perfil “pco29”, e em roxo a comunidade centrado no perfil “342artes”. A publicação em destaque do usuário “342artes” trata-se de um recorte específico, discutindo em um contexto do setor cultural, o perfil é fruto de uma campanha promovida por um grupo de artistas brasileiros, “342 Artes - Contra a censura e difamação”. o perfil “pco29” pertence ao Partido da Causa Operária, partido político brasileiro de extrema-esquerda, em sua publicação principal associa a defesa de limites à liberdade de expressão e de pensamento com a defesa à ditadura.

Relacionando as palavras-chave com maior ocorrência nas publicações coletadas, é possível identificar a grande representatividade dos termos conectados com as temáticas das principais *clusters*.

Figura 6 - Análise de palavras-chave por ocorrência

Fonte: Análise disponibilizada pelo software Netlytic.

Nesse cenário é importante destacar que no contexto das plataformas de redes sociais, a credibilidade está diretamente ligada com o engajamento, então quem obtém uma grande quantidade de curtidas, comentários e compartilhamentos, acaba sendo visto como uma autoridade por outros usuários, independente do seu aprofundamento na temática discutida, podendo até mesmo se tratar do compartilhamento de uma informação incorreta.

Fica cada vez mais evidente que a confiabilidade deixa de estar conectada ao conhecimento de um indivíduo ou reputação de uma plataforma, mas passa a ser considerada por números, desencadeando uma crise nos fundamentos da verdade. O engajamento de determinadas informações, acabam por afetar a percepção do leitor, gerando um efeito manada, desprezando-se a importância da fonte de informação. Sampaio, Lima e Oliveira (2018, p. 1666) apontam que essa crise está diretamente ligada ao “[...] esmaecimento do poder das autoridades, dos especialistas, pessoas que conhecem em profundidade determinados campos do conhecimento”, sendo alavancada pelo paradigma tecnocêntrico, que naturaliza o uso intenso de tecnologias.

Nessa dinâmica fica evidente que poucos usuários exercem uma grande influência diante dos demais

5. Considerações Finais

O presente trabalho abordou as articulações discursivas e interações ligadas à liberdade de expressão no Twitter, utilizando procedimentos de Análise de Redes Sociais, explorando as conexões entre os perfis, para depois compreender a relação entre os tópicos mais relevantes identificados nos textos das postagens. Neste contexto, foi possível identificar como alguns atores podem exercer influência na rede e nos tópicos discutidos, destacando o papel das autoridades no direcionamento das narrativas.

É importante ressaltar que algumas limitações dos softwares trabalhados dificultam a realização de uma análise mais ampla, principalmente pelo fator de limitação de período de coleta e quantidade de postagens recuperadas.

Faz-se essencial refletir a respeito da legitimação das informações, pois a própria ideia de legitimidade se altera nesse contexto das plataformas de mídias sociais, em uma dinâmica onde o compartilhamento e validação da informação está ligada a uma alta disseminação e engajamento. Esse parâmetro de legitimação contrasta com os métodos diretamente controlados por intermediários, com conhecimento especializado, responsabilidade com verificação e compartilhamento da informação, além de um compromisso ético (LIMAYE et al., 2020, p. e277).

A pesquisa em andamento permitiu compreender como a análise baseada na representação da informação através de grafos, em um contexto da Web Semântica, possibilita uma identificação das relações de influência, observando o impacto que as autoridades digitais podem exercer diante de um debate nas plataformas de mídias sociais. O estudo ressalta essa combinação

de técnicas como uma alternativa para experimentos com dados ligados da Web, possibilitando a identificação de elementos narrativos e mapeamento das relações de influência.

Referências

ARAÚJO, Gabriela; MORAES, Fabricio; PISA, Ivan. Análise exploratória de dados do Twitter: compreendendo as conexões da informaçã o de saúde durante o surto da febre am arela em 2017. ***Brazilian Journal of Inf ormation Science:*** Research trends, [S . l.], v. 14, n. 3 - jul-set, p. e020006, 2020. DOI: 10.36311/1940-1640.2020.v14n3.10179. Disponível em: https://revistas.marilia.unesp.br/index.php/bjis/article/view/10179. Acesso em: 2 dez. 2021.

CÂMARA DOS DEPUTADOS. Sessão Solene - ***Entrega do Prêmio Transparência e Fiscalização Pública*** – 07/12/2021. Youtube, 7 dez. 2021. Disponível em: https://www.youtube.com/watch?v=pUaH8Xju770&list=TLGGAr9HhBRg22cxNzAzMjAyMg. Acesso em: 17 jan. 2022.

EMPINOTTI, M. L.; PAULINO, R. DE C. R.; SERUFFO, M. C. DA R.; PIRES, Y. P.; DE SOUZA, K. E. S. Participação popular na produção e compartilhamento de informação: caso #CoronaVirusBrasil + "Bolsonaro" no Twitter. ***Rizoma***, v. 9, n. 1, 4 nov, p.153-168, 2021. DOI: 10.17058/rzm.v9i1.16436. Disponível em: https://online.unisc.br/seer/index.php/rizoma/article/view/16436. Acesso em: 2 dez. 2021.

KALINKE, Priscila; DA ROCHA, Anderson Alves; CASTANHEIRA, Karol Natasha Lourenço. Entre Bolhas: uma análise de formação de redes no Twitter no contexto da pandemia do novo Coronavirus no Brasil. ***Revista Brasileira de História da Mídia***, v. 9, n. 2, p.59-79, jul./dez. 2020. DOI: 10.26664/issn.2238-5126.92202011467. Acesso em: 2 dez. 2021.

LIMAYE, Rupali Jayant; SAUER, Molly; ALI, Joseph; BERNSTEIN, Justin; WAHL, Brian; BARNHILL, Anne; LABRIQUE, Alain. Building trust while influencing online COVID-19 content in the social media world. ***The Lancet Digital Health,*** [S. l.], v. 2, n. 6, p. e277–e278, 2020. DOI: 10.1016/S2589-7500(20)30084-4. Disponível em: <https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30084-4/fulltext>. Acesso em: 1 abr. 2022.

NUNES, Flávia Rodrigues Elias; MACULAN, Benildes Coura Moreira dos Santos; ALMEIDA, Maurício Barcellos. Os fundamentos da Web Semântica como ferramentas de auxílio para as demandas da Sociedade da Informação. **Em Questão**, Porto Alegre, v. 26, n. 3, p. 224–249, 2020. DOI: 10.19132/1808-5245263.224-249. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/92336. Acesso em: 03 nov. 2022.

RECUERO, R.; BASTOS, M.; ZAGO, G.. ***Análise de redes para mídia social.*** Porto Alegre: Sulina, 2015.

RECUERO, Raquel e GRUZD, Anatoliy. Cascatas de Fake News Políticas: um estudo de caso no Twitter. ***Galáxia***, n. 41, p. 31-47, 2019. DOI: 10.1590/1982-25542019239035. Disponível em: https://doi.org/10.1590/1982-25542019239035. Acesso em: 2 dez. 2021

RECUERO, R.; SOARES, F. O Discurso Desinformativo sobre a Cura do COVID-19 no Twitter: Estudo de caso. ***E-Compós***, [S. l.], v. 24, 2021. DOI: 10.30962/ec.2127. Disponível em: https://www.e-compos.org.br/e-compos/article/view/2127. Acesso em: 2 dez. 2021.

ROMEIRO PAULINO, Rita de Cássia e PIRES VENTURA, Mariane. O engajamento no Twitter: Métodos de análise para #Somos70porcento. ***Cuad.inf***, n.49, pp.51-71, 2021. DOI: 10.7764/cdi.49.27293. Disponível em: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0719-367X2021000200051&lng=es&nrm=iso. Acesso em: 2 dez. 2021.

VINHAS, O.; SAINZ, N.; RECUERO, R. Antagonismos discursivos nas hashtags #marqueteirosdojair e #bolsolão no Twitter nas eleições de 2018 no Brasil: contribuições da análise de redes sociais à sociologia digital. ***Estudos de Sociologia***, [S. l.], v. 25, n. 48, 2020. DOI: 10.52780/res.13433. Disponível em: https://periodicos.fclar.unesp.br/estudos/article/view/13433. Acesso em: 2 dez. 2021.

ARAÚJO, C. Pós-verdade: novo objeto de estudo para a Ciência da Informação. **Informação & Informação,** [S.l.], v. 26, n. 1, p. 94-111, mar. 2021b. ISSN 1981-8920. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/39667>. Acesso em: 05 maio 2021.

CASTELLS, M. **A Sociedade em rede: a era da Informação: economia, sociedade e cultura.** São Paulo: Paz e Terra, 2011. v. 1.

FOUCAULT, M. **Microfísica do poder.** Rio de Janeiro: Graal, 2006.

FONSECA, J. P., **Poder, Biopolítica e Governamentalidade.** Belo Horizonte, 2009. Disponível em: <https://repositorio.ufmg.br/handle/1843/VCSA-8BNR2J>. Acesso em: 13 nov. 2021.

HIGGINS, S. S., & RIBEIRO, A. C.. **Análise de redes em Ciências Sociais.** Brasília: Enap, 2018. Disponível em: <https://repositorio.enap.gov.br/bitstream/1/3337/1/Livro_Analise de Redes em Ciências Sociais.pdf> Acesso em: 10 out. 2021.

PARISER, E. (2011). **O filtro invisível: O que a internet está escondendo de você**. Nova York, Estados Unidos: The Penguin Press.

SOARES, Felipe B. Circulação de informação no twitter: como líderes de opinião ressignificam as notícias. XXIX Encontro Anual da Compós, Universidade Federal de Mato Grosso do Sul. **Anais,** Campo Grande - MS, 2020 Disponível em: <http://www.compos.org.br/biblioteca/trabalhos_arquivo_6OOX9F1O632DAU0VLTQG_30_8339_01_03_2020_09_18_41.pdf>. Acesso em: 25 jul. 2021.

METADADOS PARA COLEÇÕES E ACERVOS ARTÍSTICOS UNIVERSITÁRIOS

METADATA FOR UNIVERSITY ARTISTIC COLLECTION

Aline Cristina Gomes Ramos[1], Daniela Lucas da Silva Lemos[2]
(1) Fundação Getúlio Vargas, Vitória/ES, alinecgramos@yahoo.com
(2) Universidade Federal do Espírito Santo, Vitória/ES, daniela.l.silva@ufes.br

Resumo
O presente trabalho visa explanar parte da pesquisa realizada no mestrado em História, Política e Bens Culturais do Centro de Pesquisa e Documentação de História Contemporânea do Brasil da Fundação Getúlio Vargas, sendo parte inicial de um esquema de metadados para documentação de acervos e coleções de arte universitários brasileiros, com base no caso do Centro de Artes da Universidade Federal do Espírito Santo. Para tanto, estabelece-se o cotejamento de normas e diretrizes, comparando os campos e adotando aqueles com maior índice de repetição e alinhamento semântico à especificidade das obras. Como resultado, apresenta-se uma sistematização com o *crosswalk* realizado a partir de padrões de documentação a nível nacional e internacional, organizando uma proposta para levantamento e registro do acervo artístico universitário. Conclui-se que a recorrência de um metadado demonstra o grau de importância e forma imprescindível, sendo possível criar um modelo consistente e contemporâneo para o registro da informação em arte.

Palavras-chave: Organização da informação; Acervo artístico universitário; Crosswalk; Metadados.

Abstract
The present work aims to explain part of the research carried out in the master's degree in History, Politics and Cultural Goods of the Center for Research and Documentation of Contemporary History of Brazil of Getúlio Vargas Foundation, seeking to propose an initial part of a metadata scheme for documenting collections and collections of Brazilian university art, based on the case of the Arts Center of the Federal University of Espírito Santo. To this end, norms and guidelines are collated, comparing the fields and adopting those with the highest repetition rate and semantic alignment to the specificity of the works. As a result, a systematization is presented with the crosswalk carried out based on national and international documentation standards, organizing a proposal for the survey and registration of the university's artistic collection. It is concluded that the recurrence of a metadata demonstrates the degree of importance and essential form, making it possible to create a consistent and contemporary model for recording information in art.

Keywords: Information organization; University artistic collection; Crosswalk; Metadata.

1. Introdução

Acervo e coleção, de acordo com a Enciclopédia Itaú Cultural (2018), são utilizados recorrentemente como sinônimos nos países ibero-americanos em publicações leigas. Todavia, existem distinções apesar das inúmeras semelhanças de significados e de ambas derivarem do latim, definindo, genericamente, uma reunião ou um conjunto de coisas ou objetos que compõem patrimônio. Há uma relação hierárquica entre os termos, em que acervo costuma designar um corpo mais amplo, constituído de várias coleções de propriedade pública ou privada, enquanto coleção é um conjunto, organizado, reunido pelo valor artístico, cultural, histórico de seus componentes, ou por sua raridade, singularidade, ou pelo interesse do colecionador.

A especificidade do tema exige outra distinção, a de acervos e coleções universitárias. Torna-se necessário, dentre outros, o entendimento das funções, origens e históricos de formação dessas coleções, pois resumir o qualitativo universitário apenas ao fato de se estar sob a tutela de instituição de tal caráter leva ao reducionismo do que, em verdade, pode-se acrescentar sobre a discussão. As coleções universitárias formadas no âmbito das atividades acadêmicas, envolvidas em projetos e pesquisas, “são expressões de categorias do conhecimento e testemunhas de formas sensíveis, materiais e empíricas, de se produzir e disseminar o saber científico” (JULIÃO, 2015, p.16).

No Brasil, o processo de criação das universidades acompanha simultaneamente

a formação do patrimônio universitário de coleções, seja por meio de doação, pesquisa ou aquisição, sendo justificado por serem instituições que concentram a produção de conhecimento e que ocupam posição de prestígio, historicamente, na hierarquia social. O saber e o poder tendem a estar muitas vezes em consonância, como bem discorre Pierre Bourdieu (1989) em sua teoria sobre o "poder simbólico", refletindo, neste caso, nas universidades como lugares destacados para o colecionismo, conduzindo o surgimento de coleções, acervos e museus.

Ressaltando a importância das coleções e acervo artístico universitário, suas particularidades e as dificuldades enfrentadas na sua governança, esta proposta visa estabelecer algo inicial para a organização da informação (GILLILAND, 2016) e gestão de acervo de coleções, a partir da informação em arte.

Ao longo dos anos, muitas iniciativas são criadas com finalidade de sistematizar a informação em arte, destacando-se organizações internacionais como o *Getty Research Institute*, o Comitê Internacional dos Museus, o *Collections Trust,* e no Brasil, a Resolução Normativa Nº02, de 29 de agosto de 2014, atualizada a partir da Resolução Normativa Nº06, de 31 de agosto de 2021, ambas redigidas pelo Instituto Brasileiro de Museus (IBRAM). Instituições brasileiras de caráter cultural, muitas vezes, também geram suas próprias bases para gerenciamento de seus bens, considerando a realidade local e a necessidade de trabalho, exemplificando aqui com o Projeto Simba/Donato do Museu Nacional de Belas Artes do Rio de Janeiro, o Inventário de Proteção do Acervo Cultural de Minas Gerais do Instituto Estadual do Patrimônio Histórico e Artístico de Minas Gerais, e a catalogação da Galeria de Arte Espaço Universitário da Universidade Federal do Espírito Santo. Nota-se, no entanto, que, apesar da diversidade e de não haver restrições para composição de metadados, existem alguns que são percebidos em todos os esquemas, compartilhando semântica, definição e finalidade, seja de forma literal ou aproximada.

2. Objetivos

A pesquisa tem o intuito de produzir parte inicial de um esquema de metadados compatível com a realidade dos acervos e coleções de arte universitários brasileiros, devido à especificidade dos mesmos. Como dito anteriormente, a informação em arte é objeto de discussão de diversas instituições culturais, que estabelecem modelos possíveis de serem apropriados e customizados de acordo com realidades locais.

3. Procedimentos Metodológicos

Em relação as mais conhecidas possibilidades internacionais de organização da informação para os bens culturais, o *Getty Research Institute* concebe uma planilha comparativa a partir dos grupos de informação, intitulada *Metadata Standards Crosswalk*, tendo por referência o *Categories for the Description of Works of Art* (CDWA).

A comparação dos padrões tem por intermédio a reunião e a *expertise* dos investigadores da área, portanto, se estrutura de acordo com renomados pesquisadores que assinam a planilha. Tal constatação é relevante para o que se pretende executar neste estudo, pois o *crosswalk* demonstra quais são os metadados mais recorrentes para a organização da informação em arte nas principais normas e diretrizes internacionais.

A repetição de um metadado atesta que sua inserção em qualquer esquema a ser criado é fundamental, no entanto percebe-se que alguns padrões da *Metadata Standards Crosswalk* englobam bens do patrimônio cultural que ultrapassam os tipos do acervo artístico (como bibliográfico, arquivístico e arquitetônico), ou são vocabulários de autoridade que não contemplam de maneira significativa os casos brasileiros ou direcionados à web semântica, sendo excluídos para a proposição.

Dessa forma, somente os seguintes são selecionados na planilha: *Categories for the Description of Works of Art* (CDWA), *Cataloging Cultural Objects* (CCO), *International Committee for Documentation* (CIDOC CRM), *Dublin Core Metadata Initiative* (Dublin Core) e *Object ID*.

Em continuidade às pesquisas, cita-se a Tese de Camila Aparecida da Silva, cujo

objetivo é a concepção de um esquema de metadados para descrição de obras de arte em museus brasileiros. Silva (2020) se vale da ISO 25964 para o *crosswalk* de dois padrões presentes na *Metadata Standards Crosswalk*, o CDWA e o CIDOC CRM, e o SPECTRUM 4.0, desenvolvido pelo *Collections Trust*.

Mesmo a ISO 25964 tendo como objeto de análise os tesauros, a parte *Interoperability with other vocabuaries* apresenta contribuições para a comparação de metadados, a partir da explanação feita para a análise de um tesauro em relação a outro e do mesmo quanto a tipos de vocabulários, que podem ser sinônimos, quase-sinônimos ou muito específicos, permitindo realizar equivalências entre termos e conceitos, além da combinação de um ou mais termos para alcançar uma composição análoga. Há também ocasiões em que os metadados estão igualmente nomeados, mas exercem funções diferentes e outros, em que os metadados têm nomes diferentes, mas as mesmas finalidades. Da mesma maneira, existem situações de ausência de equivalentes perfeitos, como em casos de não coincidência total entre termos/conceitos e de lacunas. Recomenda-se também usar um termo específico na falta de um termo genérico ('equivalente aproximativo') e redigir uma nota explicativa. Outra opção é apenas indicar a falta de equivalências.

Assim, aos poucos, se evidencia que um esquema de metadados reflete uma perspectiva, pois os termos ou unidades de informação só adquirem sentido a partir da organização de um conjunto sistematizado por um grupo, com definições, hierarquias e correlações, ou seja, um sistema de conceitos, com ordem e subordinação lógica.

No momento, poder-se-ia considerar satisfatório o *crosswalk* conseguido com a planilha do Getty e a síntese de Silva, afinal resultam na comparação de sete padrões voltados para informação em arte de incontestável reconhecimento internacional. No entanto, é unânime nas orientações dessas normas e diretrizes que se almeja que a Linguagem Documentária divulgada por estes especialistas possa ser customizada localmente. Por isso, opta-se por agregar experiências desenvolvidas no Brasil por órgãos como o Instituto Brasileiro de Museus (IBRAM), o Museu Nacional de Belas Artes do Rio de Janeiro (MNBA/RJ), o Instituto Estadual do Patrimônio Histórico e Artístico de Minas Gerais (IEPHA/MG), e a da Galeria de Arte Espaço Universitário (GAEU).

Deste modo, a partir da comparação de normas e diretrizes internacionais e brasileiras, confecciona-se uma sugestão com os campos mais recorrentes, somado a outros direcionados para o contexto das universidades.

4. Resultados

Os resultados são apresentados no Apêndice A, onde se encontram a comparação dos metadados das normas e diretrizes e a síntese proposta, finalizada com trinta elementos de metadados, a saber: denominação; classificação; título/ título da série; autoria/ atribuição; data/ época; origem; dimensões; material/ técnica; marcas/ inscrições/ legendas; estado de conservação; tema; descrição; situação – localização anterior/ específica/ atual; assinatura; forma; número de patrimônio; número de inventário; mídias relacionadas; coleção; procedência; documentação fotográfica; condições de segurança; intervenções – descrição/ responsável/ onde/ data da última avaliação; data de aquisição; forma de aquisição; doador/ vendedor; valor; desincorporação e alienação – forma de alienação/ data da última avaliação; etiquetagem – etiquetável/ onde?; realização inventário.

Ao parear os metadados das diferentes fontes, percebem-se mais semelhanças do que disparidades entre eles, variando, em verdade, a quantidade e o detalhamento dos itens de acordo com a finalidade de cada norma e diretriz, somado a especificidades recomendadas para o preenchimento dos campos.

Outro detalhe importante, é que as normas e diretrizes recomendam os metadados, inclusive com a CDWA indicando os *core* (essenciais), porém são consensuais em afirmar que funcionam apenas como proposições genéricas, a serem incorporadas com base nas vivências das instituições. Por isso, estabelecem-se aqui os trinta elementos para documentação do acervo artístico

universitário, em consequência do manejo cotidiano com obras de arte da UFES, todavia outras organizações podem considerar relevantes metadados não copilados e necessários para retratar suas características.

5. Considerações Finais

De acordo com o elencado acima, considerando que as diretrizes internacionais são consideradas modelos abertos e adaptáveis para às realidades locais, trazendo também normas elaboradas no contexto brasileiro e igualmente para a estrutura universitária, é possível propor um *crosswalk* para sintetizar uma lista de metadados fundamentais para a organização da informação em arte no âmbito universitário.

Esta sistematização é uma proposta inicial de um esquema de metadados do qual se pretende em pesquisa futura formalizá-lo por meio de adequações semânticas, sintáticas, configurações de tipos de dados para cada elemento de metadados, regras de preenchimento, incluindo termos aceitáveis para cada metadado por meio de linguagens documentárias previamente selecionadas para a representação do conhecimento do domínio.

Referências

ACERVO e Coleção. *In*: ENCICLOPÉDIA Itaú Cultural de Arte e Cultura Brasileira. São Paulo: Itaú Cultural, 2022. Disponível em: http://enciclopedia.itaucultural.org.br/termo14329/acervo-e-colecao. Acesso em: 04 de maio de 2022. Verbete da Enciclopédia.

BACA, M.; HARPRING, P.; WARD, J.; BEECROFT, A.; CLARKE, S.; SILVA, C.; EKLUND, J.; GILLILAND, A. J.; O'KEEFE, E.; WOODLEY, M. S.**Metadada Standards Crosswalk**. Disponível em: https://www.getty.edu/research/publications/electronic_publications/intrometadata/crosswalks.html. Acesso em: 15 maio 2022.

BRASIL. Resolução Normativa Nº02, de 29 de agosto de 2014. Disponível em: https://www.gov.br/conarq/pt-br/legislacao-arquivistica/resolucoes/resolucao-normativa-no-2-de-29-de-agosto-de-2014. Acesso em: 24 jun. 2021.

COLLECTIONS TRUST. **SPECTRUM 4.0**: Padrão para gestão de coleções de museus do Reino Unido/ Collections Trust. São Paulo: Secretaria de Estado da Cultura; Associação de Amigos do Museu do Café; Museu de Imigração do Estado de São Paulo; Museu da Ciência da Universidade de Coimbra; Pinacoteca do Estado de São Paulo. São Paulo, 2014. Disponível em: https://spectrum-pt.org/2014/09/spectrum-4-0-versao-digital-em-portugues-ja-disponivel/. Acesso em: 20 maio 2021.

FERREZ, H. D.; PEIXOTO, M. E. S. **Manual de Catalogação**: Pintura – Escultura – Desenho – Gravura. Rio de Janeiro: Museu Nacional de Belas Artes, 1995.

GILLILAND, A. J. Setting the Stage. *In*: BACA, M. (Ed.). **Introduction to metadata.** 3. ed. Los Angeles: Getty Research Institute, 2016.

IEPHA/MG. **Inventário de Proteção do Acervo Cultural de Minas Gerais**. Belo Horizonte: Secretaria de Estado de Cultura de Minas Gerais, 2009. Disponível em: http://www.iepha.mg.gov.br/images/Documentos/Programas/MODELO_DE_FICHAS_IPACMG.pdf. Acesso em: 5 maio 2021.

JULIÃO, L. Museu e coleções universitárias. MORENO, A.; NASCIMENTO, A. (org.). **Universidade, memória e patrimônio**. Belo Horizonte: Mazza Edições, 2015. p.13-24.

SILVA, C. A. **Esquema de metadados para descrição de obras de arte em museus brasileiros**: uma proposta. Orientadora: Dra. Marilda Lara. 2020. 646 f. Tese (Doutorado em Ciência da Informação) – Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 2020. Disponível em: https://www.teses.usp.br/teses/disponiveis/27/27151/tde-01032021-162722/. Acesso em: 24 jun. 2021.

6. Apêndice A – Comparação de metadados com formulação de proposta.

CDWA	CCO	CIDOC	*Dublin Core*	*Object ID*	INBCM/ IBRAM	SPECTRUM	Simba/ Donato	GAEU	IPAC/MG	PROPOSTA
Object/ Work Type	Work Type	P41 Classified P2 has type	Type	Type of Object	Denominação	-	-	Tipo de acervo	Designação	Denominação
Classication Term	Class	E89 Propositional Object	Subject (Classification schema)	-	Classificação	Termo de classificação	Objeto	Classe	Espécie	Classificação
Title Text	Title	E35 Title P102 has title P109 has symbolic content	Title	Title	Título	Título	Título	Título	-	Título/ Título da série
Creator Description	Creator Display	E12 Production E7 Activity E65 Creation E21 Person E39 Actor E74 Group E67 Birth E69 Death	Creator	Maker	Autor	Criador	Autor	Autoria	Autoria	Autoria/ Atribuição
Creator Date	Display Date	E52 Time-Span E61 Time Primitive	Date: Created	Date or Period	Data de Produção	Data de Criação	Data	Data	Época	Data/ Época
Creation Place/ Original Location	Creation Location	-	Subject or coverage.spatial	-	Local de Produção	-		Origem	Origem	Origem
Dimensions Description	Measurements Display	E54 Dimension E16 Measurements P43 has dimension P90 has value P91 has unit	Format. Extent (schema)	Measurements	Dimensões	Medidas	Dimensões da obra	Dimensões	Dimensões	Dimensões
Materials/ Techniques Description	Material/ Technique Display	E57 Material P45 consists of P32 used general technique		Materiais e Técnicas	Material/ Técnica	Materiais/ Técnicas	Material/ Técnica	Material/ Técnica	Material/ Técnica	Material/ Técnica
Inscription Transcription or Description	Inscriptions		Description	Inscriptions and Markings			Marcada? Onde?	Inscrições/ posição	Marcas/ Inscrições/ Legendas	Marcas/ Inscrições/ Legendas
Conservation/ Treatment Description	Conservation/ Treatment History		Description	Distinguishing Features			Estado de Conservação	Estado de Conservação	Estado de Conservação	Estado de Conservação
Subject Display	Subject Display	Subject	Description Abstract	Subject	-	Assunto	Tema	-	-	Tema

General Subject Terms	Subject	E36 Visual Item P129 is about P62 depicts P138 represents P67 refers to P128 carries P65 shows visual item P190 has symbolic content	Subject or coverage.spatial or coverage.temporal						
Descriptive Note Text	Description		Description	Description	Resumo Descritivo	Descrição Formal	Descrição	Descrição	Descrição
Repository/ Geographic Location	Current Location	P53 Places has forme ror current location? [Domam: physical thing] P54 has current permanente location [Domain: physical object] P55 has current location [Domain: physical object]			Situação	Localizada?			Situação – Localização Anterior/ Específica/ Atual
						Assinada? Onde?	Assinatura/ Posição		Assinatura
			Format			Formato			Forma
					Nº de Registro	Nº de Registro	Nº de Identificação		Nº de Patrimônio
					Outros Números	Nº de Inventário	Outros Números		Nº de Inventário
					Mídias relacionadas	Mídias relacionadas			Mídias relacionadas
						Coleção/ Classe	Coleção	Acervo	Coleção
						Procedência	Procedência	Procedência	Procedência
						Foto	Foto	Documentação Fotográfica	Documentação Fotográfica
								Condições de segurança	Condições de segurança
						Restaurado?/ Data da última avaliação	Intervenções/ Restauração/ Data	Intervenções - Intervenções – Responsável/ Data	Intervenções – Descrição/ Responsável/ Onde/ Data da

					última avaliação
	Data de Aquisição	Data de Aquisição	Data de Aquisição		Data de aquisição
	Forma de Aquisição	Forma de Aquisição	Forma de Aquisição		Forma de aquisição
		Doador/ Vendedor	Doador/ Vendedor		Doador/ Vendedor
		Valor de Compra			Valor
	Desincorporação e Alienação/ Forma de Alienação/ Data da última avaliação				Desincorporação e Alienação – Forma de Alienação/ Data da última avaliação
		Título para etiqueta			Etiquetagem – Etiquetável/ Onde?
		Catalogação		Ficha Ténica	Realização inventário

Fonte: Dados da Pesquisa, 2022.

MÉTODOS ÁGEIS NA CIÊNCIA DA INFORMAÇÃO: ENSINO DO SCRUM

AGILE METHODS IN INFORMATION SCIENCE: TEACHING SCRUM

Patrícia Nascimento Silva[1]
(1) Programa de Pós-Graduação em Gestão & Organização do Conhecimento - Universidade Federal de Minas Gerais, Belo Horizonte - MG - Brasil, patricians@ufmg.br

Resumo
A competência em dados envolve desafios da gestão de projetos e da tecnologia e abarca demandas de produtos e serviços de informação que adotam processos ágeis de desenvolvimento. Com isso, conhecer e desenvolver habilidades para atuar em equipes ágeis é mais uma competência confiada ao profissional da informação. Este artigo apresenta um relato de experiência sobre a modelagem e desenvolvimento de uma disciplina optativa sobre o método Scrum, no contexto de produtos e serviços de informação, no Programa de Pós-Graduação em Gestão & Organização do Conhecimento na Escola de Ciência da Informação da Universidade Federal de Minas Gerais. A proposta da disciplina visa capacitar o discente para atuar em equipes ágeis e conhecer práticas atuais de mercado. A disciplina é amparada em dois eixos: (1) Eixo teórico e (2) Eixo aplicado e contempla uma visão geral sobre as origens das metodologias ágeis, os métodos existentes e o detalhamento do método Scrum. Como resultados parciais destacam-se o preenchimento de todas as vagas disponibilizadas na oferta do segundo semestre de 2022, a motivação e interação dos discentes nas aulas. Espera-se que ao concluir a disciplina os discentes estejam preparados para integrar equipes ágeis e futuramente, alcançarem posições de liderança nestas equipes.

Palavras-chave: Métodos ágeis; Scrum; Ciência da Informação; Profissional da Informação; Competência em dados.

Abstract
Data competence involves project management and technology challenges and encompasses demands for information products and services that adopt agile development processes. Thus, knowing and developing skills to work in agile teams is another competence entrusted to the information professional. This paper presents an experience report on the modeling and development of an optional discipline on the Scrum method, in the context of information products and services, in the Postgraduate Program in Management and Knowledge Organization at School of Information Science at the Federal University of Minas Gerais. The purpose of the discipline aims to train the student to work in agile teams and learn about current market practices. The discipline is based on two axes: (1) Theoretical axis and (2) Applied axis and includes an overview of the origins of agile methodologies, the existing methods and the detailing of the Scrum method. As partial results, we highlight the filling of all vacancies available in the offer of the second semester of 2022, the motivation and interaction of students in classes. It is expected that, upon completing the course, students are prepared to join agile teams and, in the future, reach leadership positions in these teams.

Keywords: Agile methods; Scrum; Information Science; Information Professional; Data competence.

1 Introdução

O profissional da informação tem acompanhado as evoluções tecnológicas e as novas demandas por serviços e produtos de informação. A competência em dados compreende as habilidades técnicas e tecnológicas, reconhecimento de recursos de dados, e principalmente a facilidade de comunicação oral com seus usuários e o gerenciamento de projetos (FEDERER, 2018; KOLTAY, 2019). A transformação digital em instituições e organizações impulsionou o uso de novas tecnologias digitais utilizadas para melhorar a experiência do cliente, otimizar operações e criar novos modelos de negócio (FITZGERALD et al., 2014). Acelerada pela pandemia de COVID-19, em 2020, a dimensão dos dados passou a ser observada e valorizada em projetos que ainda não compreendiam o significado e a importância dessa temática.

Os expressivos volumes de dados advindos de diversas fontes heterogêneas aumentaram a complexidade da temática de dados que integra um dos desafios da sociedade da informação. A análise massiva de dados, a criação de plataformas digitais são atividades em expansão (BERRíO-ZAPATA; RODRIGUES; GOMES, 2022) e

campos de atuação do profissional da informação.

A adoção de ciclos curtos e interativos no desenvolvimento de produtos é uma das principais características do desenvolvimento ágil (PRESSMAN, 2006) e tem sido uma das estratégias para gestão de atividades que envolvem muitas mudanças e entregas constantes.

Os métodos ágeis surgiram nos processos de desenvolvimento de *software* nos anos 1990 e foram implementados por grandes empresas a partir dos anos 2000. Atualmente a grande maioria das empresas que desenvolvem *softwares*, independente de seu tamanho ou do foco de seu negócio, usa princípios ágeis, em maior ou menor escala (VALENTE, 2020). Contudo, os métodos também têm sido utilizados em diferentes domínios como alternativa para projetos complexos envolvendo desenvolvedores, pesquisadores, analistas, cientistas e outros especialistas (SCHWABER; SUTHERLAND, 2020). Destaca-se que empresas que aderiram ao trabalho remoto ou híbrido também têm utilizado métodos ágeis para coordenar atividades realizadas pelos seus colaboradores domiciliados em diversas localidades.

A ciência da informação tem pautado novos caminhos na era baseada em dados, através dos conceitos e tecnologias que envolvem Web Semântica, Ontologia, Dados Ligados. Assim, a área entra definitivamente na rota de interesse de muitas outras áreas de conhecimento, que tem entendido que grandes partes desses estudos competem e dependem de pesquisas realizadas estritamente na Ciência da Informação, com significativo apoio da Ciência da Computação (SANTARÉM SEGUNDO, 2018). Com isso, além de compartilhar técnicas, as áreas compartilham métodos e formas de trabalho em projetos que integram diversos profissionais.

Desenvolver competências na gestão de projetos é uma habilidade esperada por um bibliotecário e pelo profissional da informação e essa foi uma das motivações para criação de uma disciplina que abordasse o método Scrum. A crescente demanda por profissionais da informação em equipes de desenvolvimento de produtos e serviços digitais faz com que competências e habilidades para integrar equipes ágeis seja habitual para este profissional. Diante da relevância do conteúdo e da necessidade de sua inserção na academia, este estudo justifica-se para compartilhar as experiências vivenciadas pela docente na proposição da disciplina.

2 Objetivos

O objetivo do artigo é apresentar a proposta desenvolvida para a disciplina Processos Ágeis de Desenvolvimento de produtos no Programa de Pós-Graduação em Gestão & Organização do Conhecimento na Escola de Ciência da Informação da Universidade Federal de Minas Gerais. A disciplina visa capacitar profissionais da informação em métodos ágeis para integrar equipes multidisciplinares em diferentes áreas do conhecimento. Para tanto foram definidos quatro objetivos específicos: (i) Identificar as publicações da área sobre a temática, (ii) Identificar as demandas de mercado associadas ao profissional da informação, (iii) Realizar um levantamento bibliográfico sobre o Scrum e (iv) Elaborar o plano de ensino e conteúdo da disciplina.

3 Procedimentos Metodológicos

O estudo apresentado é um relato de pesquisa com abordagem quantitativa e qualitativa por meio de técnicas de observação e pesquisa documental. Para concepção da disciplina foram utilizados elementos da aprendizagem baseada em problemas, ou *Problem Based Learning* (PBL). A PBL tem foco em integrar problemas reais com habilidades de aprendizagem autônoma e de trabalho em equipe, favorecendo a adaptabilidade a mudanças e, principalmente, a solução de problemas, como pensamento crítico, criativo e aperfeiçoamento contínuo (RIBEIRO, 2008), elementos que estão diretamente relacionados ao método Scrum.

Inicialmente foi realizada uma pesquisa na base de dados BRAPCI que é uma das bases de dados referenciais de artigos e de periódicos em Ciência da Informação. A busca foi realizada em julho de 2022 utilizando o descritor “scrum”, sem recorte temporal, e retornou apenas oito trabalhos. Destes trabalhos somente um estava

relacionado às competências do profissional da informação. Uma pesquisa também foi realizada no Portal Capes, utilizando a busca por assunto que consulta todas as bases indexadas no Portal, em julho de 2022, sem recorte temporal, utilizando os descritores "scrum" e "ciência da informação" combinados, para qualquer campo. Foram retornados somente quatro artigos, sendo que dois estavam fora da temática e os outros dois já tinham sido listados na busca da base BRAPCI. Ao pesquisar por "scrum" e "biblioteconomia" combinados não foi retornado nenhum resultado.

Para identificar as demandas de mercado foi realizada uma busca na rede social LinkedIn, em agosto de 2022. Utilizou-se do recurso de busca simples na opção "vagas". Ao buscar utilizando o descritor "bibliotecário" observou-se que algumas vagas indicavam além de bibliotecário os cargos: *Product Owner*, *Data Product Owner*, *Data Librarian*, *Product Manager*. Ao pesquisar por "cientista da informação" as vagas estavam relacionadas aos cargos: *Data Scientist*, *Research Scientist*, cientista de dados. Observou-se também que a descrição das vagas, principalmente as vagas relacionadas aos cargos com nomenclatura de papéis ágeis, indicavam o conhecimento em "práticas ágeis" e em "cultura ágil". É importante destacar que foi utilizado um acesso gratuito à plataforma LinkedIn e não foi possível buscar o histórico de vagas já preenchidas ou disponíveis para candidatura ao longo dos anos, limitando a pesquisa ao momento atual da consulta.

O levantamento bibliográfico sobre o Scrum envolveu obras que abordavam a gestão de projetos e principalmente a engenharia de software, área em que o método foi criado originalmente. Publicações específicas sobre a temática como o Scrum Guide e obras popularmente comercializadas em grandes *marketplaces* como Amazon e Americanas também foram consideradas. Estes três parâmetros orientaram a seleção da bibliografia que também prezou pela experiência técnica da autora na temática. Após selecionar as referências, uma leitura analítica foi realizada em busca das principais características do método e formas de aplicação. A formulação da disciplina considerou as definições do método Scrum e aplicações no contexto da ciência da informação. Na seção resultados são apresentadas as reflexões da autora e a proposta desenvolvida para a disciplina.

4 Resultados

No contexto da ciência da informação, as buscas realizadas nas bases de dados sobre a temática Scrum, retornaram poucos documentos, demonstrando que o método ainda é pouco difundido em projetos e estudos da área. Contudo, a pesquisa de mercado realçou que várias empresas já utilizam papéis do Scrum para cargos e atividades que indicam o bibliotecário como formação / escolaridade desejada. Estes resultados evidenciam que muitas empresas já reconhecem a importância da biblioteconomia e da ciência da informação em seus domínios de atuação e incluem nomenclaturas e termos das metodologias ágeis em seus processos de trabalho e inclusive na descrição do cargo a ser ocupado pelo profissional da informação.

Com isso, optou-se por modelar uma disciplina optativa com conteúdo variável, para permitir sua constante atualização, e carga horária de 30 horas. A ementa perpassa a origem e visão geral das metodologias ágeis finalizando com um maior detalhamento sobre o método Scrum. A disciplina inicialmente foi planejada para o ensino presencial, com atividades condensadas em uma semana (15 horas) e atividades espaçadas (15 horas), possibilitando que profissionais e discentes consigam adequar seus horários para participar da atividade acadêmica.

O plano de ensino da disciplina abrangeu discussões conceituais com suporte para atividades aplicadas intercaladas, conforme conteúdo programático e atividades avaliativas planejadas, sendo estruturado em dois eixos: (1) Eixo teórico e (2) Eixo aplicado, conforme detalhado nas seções 4.1 e 4.2.

4.1 Eixo Teórico

O eixo teórico centrou-se em uma breve fundamentação teórica, resgatando as bases e princípios dos processos de desenvolvimento de software e suas

evoluções até a apresentação dos métodos ágeis, características e definições.

O desenvolvimento ágil surgiu para flexibilizar o desenvolvimento de software que envolvia grandes e complexos sistemas. As principais características desse novo processo envolviam a mudança de requisitos, advindas das pressões externas e alterações de prioridade nas organizações, e a entrega incremental, onde incrementos são entregues ao cliente para comentários e experimentação (SOMMERVILLE, 2011). Os princípios dos métodos ágeis foram utilizados em diversas propostas resultando em vários métodos como o Extreme Programming (XP), o Kanban e o Scrum que são abordados separadamente nas aulas. O método Scrum é atualmente um dos mais utilizados e a disciplina perpassa por todos os elementos que o compõem fazendo referência ao guia do método, o Scrum Guide, ou Guia do Scrum.

O Guia do Scrum é um documento criado pelos fundadores do método, Ken Schwaber e Jeff Sutherland, que contém a definição completa do método. O Guia foi traduzido para mais de 50 idiomas e as versões atualizadas são publicadas no site do projeto[1]. A última versão foi publicada em novembro de 2020 e pode ser baixada gratuitamente (SCRUM GUIDES, 2022).

Conforme definido no Guia, o framework Scrum é simples e propositalmente incompleto, apenas definindo as partes necessárias para implementar a teoria Scrum. A inteligência coletiva das pessoas é que irá construir o método, determinando sua filosofia, teoria e estrutura para atingir objetivos e criar valor. Em vez de fornecer às pessoas instruções detalhadas, as regras do Guia do Scrum irão orientar os relacionamentos e interações (SCHWABER; SUTHERLAND, 2020).

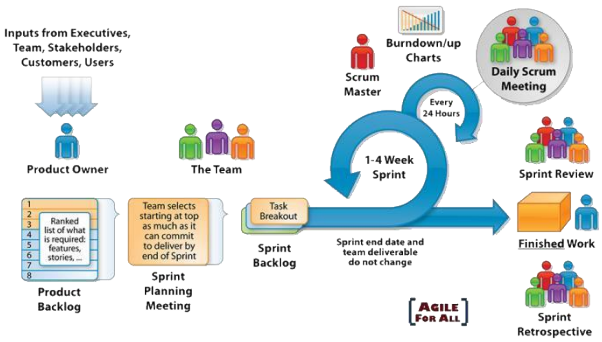
Ainda que existam definições e nomenclaturas no método Scrum, a possibilidade de adaptação é um dos pilares que deve ser sempre observado, especialmente pelo profissional da informação que poderá atuar em diferentes contextos e áreas de negócio. Desta forma, a construção teórica abordada na disciplina permite que os discentes visualizem a flexibilidade e adequação do método e, principalmente, a transparência dos processos para todos os atores envolvidos. Bibliografias que não possuem foco acadêmico também foram brevemente abordadas com o objetivo de apresentar outras fontes e discussões de outros autores sobre a temática.

Como resultado do eixo teórico foi produzido o material didático das aulas expositivas e atividades avaliativas no formato de exercícios para fixação do conteúdo.

4.2 Eixo Aplicado

O eixo aplicado centrou-se em atividades práticas e interativas da turma simulando papéis e cerimônias do método Scrum (Figura 1) no contexto de projetos da Ciência da Informação. Para tanto, inicialmente, foram elaboradas atividades individuais para fixação dos conceitos em projetos individuais, abordando questões como o *Backlog* do produto, o refinamento de requisitos e a priorização de itens. Em um segundo momento os discentes foram incentivados a trabalharem em grupos, discutindo os papéis propostos pelo método e simulando as cerimônias como, por exemplo, a reunião de planejamento da S*print*, a reunião diária e a reunião de revisão da S*print*.

Figura 1: Ciclo de Vida Scrum



Fonte: (BRAS, 2011).

Foram elaborados casos práticos que envolviam problemas de gestão e organização de dados e informações em instituições e também o desenvolvimento de produtos digitais, no contexto da ciência de dados, todos abordando problemas reais e atuais. A metodologia PBL foi uma das metodologias ativas selecionadas para a

[1] https://scrumguides.org/index.html

construção e orientação da disciplina, justamente por abordar problemas reais onde os discentes podem visualizar com mais clareza onde aplicar a teoria. Além disso, a metodologia empreende a autonomia do aluno na resolução de problemas e o pensamento crítico. O acompanhamento do percurso dos discentes, com feedbacks rápidos e o contato direto, é uma proposta para gerar uma maior motivação dos estudantes e aumentar a interação com a professora e a turma.

Um dos *cases* criados para disciplina envolve a implantação de uma solução para curadoria digital para gestão de contratos em uma organização. Nesta atividade os discentes precisam rever conceitos da Ciência da Informação, formular o ciclo de vida dos dados e analisar ferramentas tecnológicas para adequar a solução, tudo isso em consonância com os princípios do Scrum envolvendo o "cliente" e apresentando resultados para cada uma das etapas do método.

A apresentação de ferramentas para subsidiar a implantação do método Scrum foi empreendida ao longo das aulas expositivas, mas as indicações orientaram-se como sugestões. Diante da adaptabilidade do método foi estimulado durante as práticas que os estudantes identificassem ferramentas que pudessem ser utilizadas no Scrum, incluindo desde uma planilha eletrônica até *softwares* específicos para gestão de projetos e controle de tarefas.

5 Considerações Finais

A proposta modelada para a disciplina Processos Ágeis de Desenvolvimento de Produtos foi implementada na oferta do segundo semestre de 2022 e ainda está em andamento. Ao planejar a disciplina, a docente apropriou-se, principalmente, dos conceitos das metodologias ativas e das metodologias ágeis, a fim de tentar estabelecer instrumentos que permitissem utilizar estes conceitos de ambos os métodos relacionados respectivamente ao ensino e a gestão de projetos, experimentando novas expectativas no processo de ensino-aprendizagem.

Como resultado parcial pode-se destacar a grande procura dos discentes no período de matrículas e o preenchimento de todas as vagas destinadas para a disciplina. Alguns estudantes relataram que a procura foi impulsionada por acompanhar ofertas de mercado que utilizam vários termos e conceitos das metodologias ágeis e outros por trabalharem em projetos na área de tecnologia. No eixo teórico os discentes se mostraram curiosos e interessados em saber as origens e definições do método. No eixo aplicado os discentes se mantêm motivados e interagindo com toda a turma. O emprego de cenários e exemplos reais suscitam comentários sobre situações vivenciadas pelos discentes em experiências profissionais e discussões sobre o posicionamento do profissional da informação e do bibliotecário frente às novas demandas e competências esperadas pelo mercado.

Ao final do semestre será realizado um seminário onde os estudantes poderão discutir sobre as competências desenvolvidas ao longo da disciplina e apresentar sugestões para as próximas ofertas. No entanto, por meio das atividades propostas ao longo da disciplina, foi perceptível que o método Scrum requer processos bem definidos e adaptados ao contexto aplicado. Para isso, almeja-se que o profissional da informação possua competências relacionadas à análise e modelagem de processos, à gestão de projetos e conheça as principais ferramentas destas áreas. Além disso, o método exige conformidade e transparência por parte da equipe para seguir as cerimônias e executar as atividades conforme os papéis assumidos. Comunicação, desenvoltura e negociação também são indispensáveis a estes profissionais. Espera-se que ao concluir a disciplina os discentes estejam preparados para integrar equipes ágeis e futuramente, alcançarem posições de liderança nestas equipes.

Referências

BRAS, Alan. Introdução ao Scrum, Alan Bras – Mestre IC Unicamp – Pesquisador em Engenharia Software Ágil (IBM), 2011. Disponível em: http://www.alanbraz.com.br/ic/scrum.pdf. Acesso em: 25 jul. 2015.

BERRíO-ZAPATA, Cristian.; RODRIGUES, Andreia Cristina Paixão; GOMES, Layane Rayssa Gaia. Plataformas, plataformização e ecossistemas de software nas bases de dados acadêmicas: aspectos conceituais. 2019. p. 361-371. Disponível em: http://hdl.handle.net/20.500.11959/brapci/125315. Acesso em: 30 ago. 2022.

FEDERER, L. Defining data librarianship: a surveyofcompetencies, skills, and training. **Journal of the Medical Library Association**, Chicago, IL, v. 106, n. 3, p. 294-303, 2018.

FITZGERALD, Michael. et al. Embracing digital technology: A new strategic imperative. **MIT sloan management review**, v. 55, n. 2, p. 1, 2014.

KOLTAY, T. Accepted and emerging roles of academic libraries in supporting research 2.0. **The Journal of Academic Librarianship**, Amsterdam, NL, v. 45, n. 2, p. 75-80, 2019.

PRESSMAN, Roger. S. **Engenharia de software**. 6. Ed. São Paulo: Makron Books, 2006.

RIBEIRO, Luis Roberto Camargo. **Aprendizado baseado em problemas**. São Carlos: UFSCAR; Fundação de Apoio Institucional, 2008.

SANTARÉM SEGUNDO, J. E. Web Semântica: fluxo para publicação de dados abertos e ligados. **Informação em Pauta**, v. 3, n. especial, p. 117-140, 26 nov. 2018. Disponível em: http://www.periodicos.ufc.br/informacaoempauta/article/view/39721. Acesso em: 19 de ago. de 2022.

SCHWABER, Ken.; SUTHERLAND, Jeff. **The scrum guide-the definitive guide to scrum**: The rules of the game. SCRUM.org, Nov-2020, 2020. Disponível em: https://scrumguides.org/docs/scrumguide/v2020/2020-Scrum-Guide-PortugueseBR-3.0.pdf. Acesso em: 19 abril 2022.

SCRUM GUIDES. The 2020 Scrum Guide. 2022. Disponível em:https://scrumguides.org/. Acesso em: 10 ago. 2022.

SOMMERVILLE, Ian. **Engenharia de Software**. 9 ed. São Paulo: Pearson Prentice Hall, 2011.

VALENTE, Marco Túlio. **Engenharia de Software Moderna**: Princípios e Práticas para Desenvolvimento de Software com Produtividade, Editora: Independente, 395 páginas, 2020.

MODELO DE PRESERVAÇÃO HIPATIA: METODOLOGIA DE ESTUDO DE METADADOS PARA EXTRAÇÃO

HIPATIA PRESERVATION MODEL: STUDY CASE OF THE METADATA EXTRACTION METHODOLOGY

Ívina Flores Melo[(1)], Tatiana Canelhas[(2)], Tiago Emmanuel Nunes Braga[(3)]
(1) Instituto Brasileiro de Informação em Ciência e Tecnologia, ivinamelo@ibict.br
(2) Instituto Brasileiro de Informação em Ciência e Tecnologia, tatianacanelhas@ibict.br
(3) Instituto Brasileiro de Informação em Ciência e Tecnologia,tiagobraga@ibict.br

Resumo
Trata-se de trabalho qualitativo descritivo que tem por objetivo descrever, brevemente, a metodologia de estudo de metadados do Modelo de Preservação Hipatia. O modelo de preservação Hipatia, desenvolvido e pesquisado pelo Instituto Brasileiro de Informação em Ciência e Tecnologia aplica o Modelo OAIS em ambientes digitais institucionais e é uma camada de barramento tecnológico que automatiza o processo de preservação digital. Como objetivos específicos pretende-se relatar o processo de extração de dados executado nos projetos do modelo de preservação Hipatia e apresentar um case (Case Arquivo Nacional). O relato do processo de extração dá-se pela explicação das sete etapas necessárias para a elaboração do dicionário de dados. O relato do estudo de caso demonstra a aplicação das etapas de estudo de metadados. A partir do modelo de referência OAIS, empreende-se esforços na formatação de ambientes sistêmicos e interoperáveis de maneira a viabilizar a preservação digital. Por meio do Hipatia, é possível vislumbrar que o OAIS possa ganhar espaços e amadurecer o cenário brasileiro na temática.

Palavras-chave: Modelo de Preservação Hipatia. Extração de dados. Preservação Digital.

Abstract
This is a qualitative research that aims to describe the metadata study methodology used by the Hipátia preservation model. The Hipátia preservation model is a research project developed by the Instituto Brasileiro de Informação em Ciência e Tecnologia and is based on the OAIS reference model on digital institutional environments. and developed and researched by the Brazilian Institute of Information in Science and Technology, applies the OAIS Model in institutional digital environments and it is a technological layer that automates the digital preservation process. As specific objectives, we intend to report the data extraction process performed in the projects of the Hipatia preservation model and present a case (National Archive Case). The extraction process is made by the seven steps which are necessary to write down a data dictionary. The case report demonstrated the application of the metadata study stages. Based on the OAIS reference model, efforts are made to format systemic and interoperable environments in order to enable digital preservation. Through Hipatia, it is possible to envision that the OAIS can be used by the Brazilian scenario.

Keywords: Hipatia Preservation Model. Data extraction. Digital Preservation.

1. Introdução

Na atualidade, nota-se a crescente e exponencial produção de informação e de desinformação em meios digitais. Neste universo, observa-se que a volumetria de tais informações e, consequentemente, dados produzidos em sistemas da informação demandam tratamento adequado de maneira a torná-los acessíveis ao longo do tempo e, sobretudo, autênticos e confiáveis. Este processo de tratamento, no geral, tem seus pilares no modelo de referência *Open Archival Information System* (OAIS) publicado pela ISO 14.721 (CCSDS, 2012). A exemplo disto, cita-se estratégias internacionais que se apropriam do modelo OAIS tal como o *Digital Curation Life Cycle* do Digital Curation Centre (DCC), situado no Reino Unido, o *Institutional Repository model* estudado pela Universidade de Southampton, também no Reino Unido, e o *Comprehensive Digital Preservation Services* (CDPS) liderado pelo *Massachusetts Institute of Technology* (MIT), dos Estados Unidos da América.

De maneira semelhante, no Brasil tem-se utilizado o mesmo modelo de referência (OAIS) visando a preservação digital. (SARAMAGO, 2004, LIMA et al, 2012) Neste caminho, as instituições trabalham

arduamente na elaboração de estratégias de preservação digital que contemplem os pressupostos do modelo OAIS, o que necessariamente resulta na utilização de *software* de empacotamento, repositórios confiáveis e plataforma de acesso. Este ecossistema tem em seu *core* padrões de empacotamento, tais como o METS, o PREMIS e o EAD, que orientam a normalização dos processos necessários à confiabilidade e autenticidade dos objetos preservados. Porém, esta não é uma tarefa simples. Atender à preservação digital do início ao fim, da gênese ao acesso, cumprindo todos os requisitos técnicos e tecnológicos, demonstra-se uma tarefa de grande empenho.

O modelo Hipatia, proposto pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), foi concebido a partir de uma parceria técnica iniciada em 2018 com o Tribunal de Justiça do Distrito Federal e Territórios (TJDFT) e tinha como objetivo definir uma camada de barramento tecnológico interoperável para garantir a segurança e o acesso aos documentos digitais. Da mesma forma, buscava automatizar o processo de preservação digital. Com o avançar dos estudos, a pesquisa foi ampliada para a proposição de um modelo que visasse a operacionalização e aplicação do Modelo OAIS em ambientes digitais institucionais, proporcionando aplicação técnica e viabilizando o uso de tecnologias na preservação digital. Este modelo foi delineado no Hipatia.

A aplicação do modelo Hipatia é estruturada em cinco fases: preparação arquivística, preparação computacional, extração e empacotamento de objetos digitais, preservação e disseminação. (BRAGA, 2022). Resumidamente, a preparação arquivística diz respeito à análise do sistema produtor e os objetos informacionais produzidos. Esta primeira etapa serve como suporte para a etapa de extração e empacotamento dos objetos digitais, quando é elaborada uma proposta de extração e de organização de objetos, dados e metadados. A preparação computacional baseia-se na análise de duas perspectivas, a primeira, o sistema produtor de objetos informacionais digitais, a segunda, a infraestrutura de rede da instituição. Nesta etapa são estabelecidas as diretrizes para a instalação dos software propostos pelo modelo de preservação, bem como sua configuração. Como resultado desta etapa é proposto um modelo de extração dos dados e metadados e a organização e configuração dos ambientes a fim de se manter a segurança necessária para aplicação do modelo OAIS.

Na etapa de extração de metadados são realizadas as conexões entre o ambiente de produção de objetos digitais e o ambiente de preservação. (BRAGA, 2022). Os objetos e metadados são extraídos do sistema produtor, processados, organizados em um pacote do tipo Bagit e então encaminhados para a etapa de preservação. Na interlocução entre as etapas, o estudo dos metadados do sistema de origem, ou sistema produtor, é colocado como recurso transversal, sendo contemplado pelas preparações arquivísticas e computacionais e caracterizando a extração de dados, bem como subsidiando o envio para preservação.

Sendo assim, este trabalho tem por **objetivo** descrever a metodologia proposta pelo modelo Hipatia para se fazer os estudos de metadados. Como objetivos específicos, pretende-se relatar o processo de extração de dados executado nos projetos do modelo Hipatia e apresentar o caso da pesquisa ocorrida em parceria com o Arquivo Nacional.

Este trabalho caracteriza-se por ter uma abordagem descritiva e qualitativa, bem como traz aspectos de estudo de caso.

2. Metodologia de Estudo

A conjugação das cinco etapas do modelo Hipatia resultam na extração do objeto digital e seus metadados que são enviados para o ambiente de preservação, processados em pacotes com base no Modelo OAIS. Este procedimento objetiva a preservação dos objetos digitais por longos prazos, garantindo-se sua integridade, autenticidade e segurança jurídica. O procedimento de extração, todavia, demanda que os responsáveis envolvidos tenham conhecimento aprofundado dos sistemas produtores destes objetos digitais. Tal conhecimento perpassa pelo escopo do sistema, suas funcionalidades, a arquitetura do sistema, a arquitetura do banco de dados

e os aspectos relacionados ao armazenamento dos objetos e seus formatos resultantes do uso do sistema.

Para se obter esta visão ampliada do sistema é necessário disponibilizar os acessos completos à equipe responsável pela implementação do modelo Hipatia. Não obstante, é essencial que se tenha acesso livre à toda a infraestrutura computacional, como códigos, banco de dados e diretórios de arquivos. Estes acessos são utilizados para se entender a estrutura do sistema e suas funcionalidades, bem como a dinâmica de armazenamento dos dados. É importante observar que neste momento a documentação do sistema pode ser também fonte de consulta, assim como entrevistas com gestores negociais e desenvolvedores.

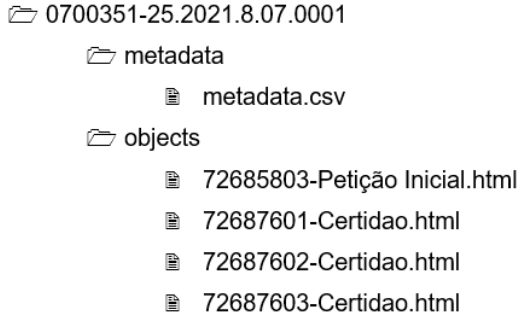
Uma vez realizado o acesso ao sistema produtor, destaca-se o escopo do sistema e inicia-se uma pesquisa em dados, informações e documentos decorrentes do escopo. Procura-se, neste momento, mapear todo o ciclo de vida da produção informacional. Este mapeamento é realizado tanto do ponto de vista da tecnologia da informação quanto do ponto de vista do usuário que busca identificar a proveniência, os contextos e as origens dos metadados descritivos.

A próxima etapa é a de extração de dados, na qual se mapeia a localização dos dados, informações e documentos em servidores locais, também conhecidos como *Local Filesystems*, em *Network Filesystems* (NFS) ou até codificados em banco de dados, do tipo *blob*, por exemplo. Nessa fase, é necessário manifestar o dado, a informação ou o documento tal qual é apresentado no sistema produtor, com todos os seus elementos internos e externos, compostos por formato, conteúdo, metadados descritivos, assinaturas, dentre outros.

Por fim, elabora-se um documento completo, intitulado Dicionário de Dados, para que a equipe de desenvolvedores possa programar no barramento que é utilizado durante a etapa de extração e empacotamento, o BarraPres. O BarraPres realiza a integração entre o sistema produtor e o ambiente de preservação. Ele executa a extração, preparo, organização de pacotes iniciais e transposição de um ambiente para outro, bem como mantém preservadas informações relacionadas ao funcionamento do modelo, tais como *logs* de ações e registros em banco.

Uma vez iniciada a extração, ela é organizada pelo BarraPres em um padrão identificável pelo sistema de preservação, com uso dos diretórios *metadata* e *objects*.

Figura 1: Dados extraídos e sistematizados pelo BarraPres



Fonte: elaboração própria

Após a extração, o Barrapres migra a pasta que foi salva com os dados que serão preservados para o formato *BagIt*[1]. Este pacote gerado é denominado Pacote de Transferência Inicial (PTI) e é apresentado na Figura 2.

Figura 2: Pacote de Transferência Inicial

[1] *BagIt* é uma padrão convencionado pela *Library of Congress* para empacotar diretórios de arquivos gerando e registrando checksums para cada arquivo armazenado em uma bag, possibilitando a verificação da integridade dos arquivos. Uma bag é o nome do file system directory que conterá minimamente data, manifest file com MD5 checksum e um bagit.txt.

Fonte: elaboração própria

Após a formatação do PTI, o BarraPres utiliza uma API-Rest para enviar o pacote ao sistema de gerenciamento do empacotamento de repositório Archivematica, de forma sistêmica quando se inicia o empacotamento e a formação do *Submission Information Package* (SIP) preconizado pelo Modelo OAIS. A API utilizada é a */api/v2beta/package*[2], que é utilizada pelo BarraPres por ser a única que possui o parâmetro de endereçamento do sistema de acesso denominado *access_system_id*. Na figura 3, segue exemplo da lista de parâmetros dessa API. O parâmetro *path* é o caminho onde está mapeado o PTI que foi convertido para base64 anteriormente.

Figura 3: Parâmetros da API de transferência

Body raw (json)

json


```
{
  "name": "0700351-25.2021.8.07.0001",
  "path": "L2hvbWUvdWJ1bnR1LzA3MDAzNTEtMjUuMjAyMS44LjA3LjAwMDE
  "type": "zipped bag",
  "processing_config": "automated",
  "accession": "",
  "access_system_id": "slug do AtoM",
  "auto_approve": true
}
```


Fonte: elaboração própria

A partir do que foi descrito nesta seção e em síntese, o estudo para extração se apresenta nas seguintes etapas:

1. Acesso ao sistema produtor e sua documentação, se for o caso;
2. Estudo do escopo e da produção informacional;
3. Mapeamento da localização de dados, informações e documentos;
4. Elaboração do dicionário de dados;
5. Configuração do BarraPres;
6. Extração de dados e criação do PTI
7. Envio do PTI para o ambiente de Preservação.

3. Estudo de caso do Arquivo Nacional

Em 2021, o IBICT firmou parceria técnica com Arquivo Nacional (AN), que teve em seu escopo principal a busca pela preservação de processos produzidos no Sistema Eletrônico de Informações (SEI). No âmbito metodológico, o estudo foi feito em um ambiente de homologação criado pelo AN somente para uso do IBICT. Selecionou-se alguns processos como exemplo. Com os números de processos, denominados NUP, em mãos, foi feita a consulta dos metadados do processo e seus objetos, documentos pertencentes ao processo, na interface de usuário do SEI. Nesta primeira consulta, os principais metadados são manifestados nas telas, como dados do processo e toda sua movimentação. O sistema utilizado pelo usuário final foi objeto de primeiro estudo para conhecimento da criação documental e todos os seus trâmites do ciclo de vida.

Figura 4: Exemplo de consulta ao banco dos interessados do documento

```
SELECT distinct c.sigla,
   c.nome,
   case
      p.sta_participacao
      when 'I' then 'Interessado'
      when 'D' then 'Destinatário'
      when 'R' then 'Remetente'
      when 'A' then 'Acesso Externo'
      else p.sta_participacao
   end as "destinatario"
FROM participante p
   INNER JOIN contato c on p.id_contato = c.id_contato
WHERE sta_participacao = 'I'
   and ID_PROTOCOLO in (
      SELECT p2.id_protocolo
      FROM protocolo p
         inner join rel_protocolo_protocolo rpp on p.id_protocolo =
               rpp.id_protocolo_1
         inner join protocolo p2 on rpp.id_protocolo_2 = p2.id_protocolo
      WHERE p.id_protocolo = id_protocolo //recuperado em consulta anterior
   );
```

Fonte: elaboração própria

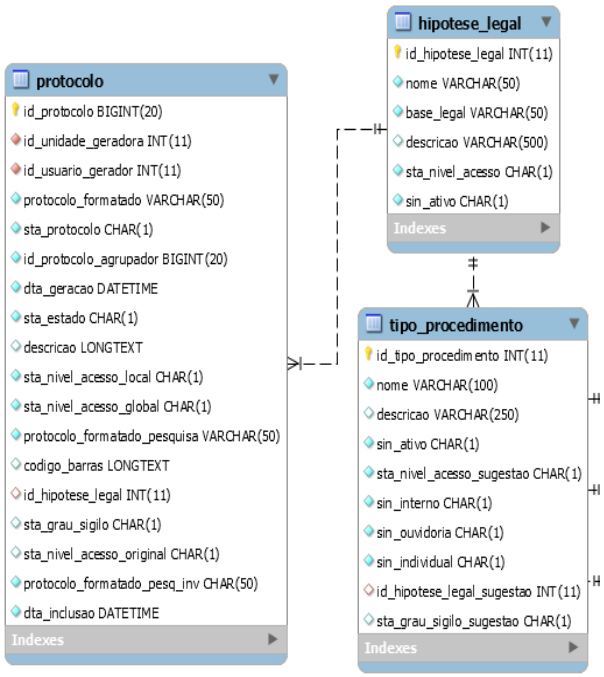
Em seguida, mapeou-se os metadados encontrados nas interfaces e buscou-se quais retornavam pelas webservices (WB) do SEI. Para aqueles metadados que não eram recuperáveis por WB, usou-se o acesso direto ao banco de dados, consultando dados como número do processo, interessados, assunto, nível de acesso, tipo de processo do SEI. Para isso foi utilizado o MySQL Workbench 8.0 CE.

[2] Disponível em https://www.archivematica.org/en/docs/archivematica-1.13/dev-manual/api/api-reference-archivematica/#package

Além dos metadados, os documentos também devem ser recuperados tal como são manifestados em tela para o usuário, em sua formatação e visualização dos principais elementos como dados da assinatura digital. Durante o estudo, percebeu-se que os documentos produzidos pelo sistema SEI são códigos *.html* salvos em diferentes campos e tabelas do banco de dados, mas todo seu conteúdo pode ser recuperado em um só campo retornado pela WB "ConsultaDocumento.LinkAcesso"[3]. Já os arquivos que são anexados aos processos ficam salvos em um *filesystem* local, mas sua recuperação acontece da mesma forma que os documentos citados anteriormente, ou seja, utilizando-se um WB fornecido pelo SEI.

Além das pesquisas diretas nas interfaces do sistema, no banco de dados e nas WS, outro método de acesso do sistema e mapeamento dos metadados deu-se apoiado por meio do Modelo de Entidade Relacionamento (MER) do SEI, possibilitando mapear outras informações tal como é apresentada na figura 3.

Figura 3- Exemplo de MER localizada



Fonte: elaboração própria

Com os metadados de gestão mapeados, comparou-se o que pode ser recuperado do SEI com os metadados citados pelo e-Arq Brasil, modelo de requisitos para sistemas de gestão arquivística de documentos, publicado pelo Conselho Nacional de Arquivos (Conarq) para órgão do Poder Executivo Federal. Coube ao AN escolher quais destes metadados seriam preservados, mesmo que eles não fossem citados pelo modelo. Aproveitou-se essa seleção também para compor alguns metadados descritivos no padrão de interoperabilidade *Dublin Core* (DC), utilizado entre o *Archivematica*[4] e o *AtoM*[5]. O formato usado para salvar os metadados no projeto do Arquivo Nacional foi o *Comma-separetad Value* (CSV), respeitando as regras de disponibilização dos dados conforme documentação do *Archivematica*. O primeiro campo deve, obrigatoriamente, ser nomeado como *filesystem*. Os seguintes foram os campos DC, seguidos dos metadados de gestão. O documento CSV foi salvo em codificação UTF-8 e teve seus valores separados por vírgulas.

Com o estudo finalizado, consolidou-se toda a informação em um documento voltado para a equipe de desenvolvimento do projeto, o Dicionário de Dados, que descrevia quais os metadados e os documentos seriam preservados, como extraí-los do SEI e como seriam disponibilizados em diretórios específicos para serem salvos no formato *BagIt*.

4. Considerações Finais

A partir do modelo de referência OAIS, empreendeu-se esforços na formatação de ambientes sistêmicos e interoperáveis de maneira a viabilizar a preservação digital. Um processo complexo e cujas demandas

[3] https://www.gov.br/economia/pt-br/assuntos/processo-eletronico-nacional/arquivos/documentacao-do-sei/sei-webservices-v3-1.pdf

[4] Archivematica é um software de empacotamento e preservação segundo o Modelo OAIS e gerenciamento de repositório confiável. É desenvolvido pela Artefectual.

[5] AToM (acesso to memory) é uma plataforma de acesso e difusão, aderente ao Modelo OAIS, desenvolvido pela Artefectual e cujo escopo debruça-se na disponibilização de representantes/derivadas de documentos preservados.

exigiram estudos que viabilizassem a aplicação do modelo de referência OAIS.

O modelo de preservação Hipatia, proposto pelo Instituto Brasileiro de Informação em Ciência e Tecnologia, aplica o Modelo OAIS em ambientes digitais institucionais e busca mediar os processos de preservação digital, tendo suas bases nas melhores práticas em cinco etapas.

Pode-se observar que o conhecimento do sistema produtor é essencial ao uso de tecnologias que viabilizem a preservação digital Foi também visto que ela apenas se consolidará com a estruturação de um ecossistema integrado mediado por padrões e normas. Por meio do Hipatia, foi possível vislumbrar que o Modelo de referência OAIS pode se tornar realidade corrente nos cenários informacionais do Brasil e se apresenta como uma solução para atender a demanda latente de implementação de Repositórios Arquivísticos Digitais Confiáveis.

Referências

BRAGA, Tiago E N. O modelo Hipatia: a proposta do IBICT para a preservação digital arquivística. **In: Hipatia: modelo de preservação para repositórios arquivísticos digitais confiáveis.** Brasília: Ibict, 2022. Disponível em: <http://labcotec.ibict.br/omp/index.php/edcotec/catalog/book/livroHipatia>. Acesso em: 30 ago 2022.

CCSDS – Consultative Committee for Space Data Systems. **Reference Model for an Open archive Information System (OAIS).** Washington: CCSDS Secretariat, June 2012. Disponível em: . Acesso em: 09 set 2022.

CONSELHO NACIONAL DE ARQUIVOS. **Modelo de requisitos para sistemas informatizados de gestão arquivística de documentos: e-ARQ Brasil.** Versão 2. Rio de Janeiro: Arquivo Nacional, 2022. Disponível em: <https://www.gov.br/conarq/pt-br/centrais-de-conteudo/publicacoes/EARQV203MAI2022.pdf>. Acesso em 09 set 2022.

LIMA, Rodrigues de Souza, A. H., OLIVEIRA, A. F., D'AVILA, R. T., CHAVES, E. P. da S. S. (2014). O modelo de referência OAIS e a preservação digital distribuída. **Ciência Da Informação, 41**(1). https://doi.org/10.18225/ci.inf.v41i1.1352 Acesso em: 07 set 2022

SARAMAGO, Maria Lurdes. Metadados para preservação digital e aplicação do modelo OAIS. In: **Actas do congresso nacional de bibliotecários, arquivistas e documentalistas.** 2004.

ORGANIZAÇÃO DE OBJETOS DE APRENDIZAGEM COM BASE NO PADRÃO LOM-IEEE UTILIZANDO DADOS LIGADOS INTEROPERÁVEIS NA WEB SEMÂNTICA

ORGANIZATION OF LEARNING OBJECTS BASED ON LOM-IEEE STANDARD USING INTEROPERABLE LINKED DATA ON THE SEMANTIC WEB

Viviane Bessa Lopes Alvarenga[1], Henrique Monteiro Cristovão[2]
(1) Universidade Federal do Espírito Santo, Vitória-ES, viviane_b_lopes@hotmail.com
(2) Universidade Federal do Espírito Santo, Vitória-ES, henrique.cristovao@ufes.br

Resumo
Devido ao aumento de cursos na Educação a Distância, tem havido um aumento na oferta de Objetos de Aprendizagem (OA) que normalmente estão disponíveis em repositórios sem interoperabilidade com outras bases, o que gera redundância de conteúdo e eleva desnecessariamente o custo de criação e manutenção, indo contrário à sua principal característica, que é a reusabilidade. Neste contexto, este trabalho propõe a investigação e seleção de um padrão de metadados para representar OA, e também a investigação e análise das peculiaridades de alguns OA disponíveis em repositórios selecionados fazendo um levantamento dos elementos necessários para a sua representação no padrão de metadados escolhido. Como esta pesquisa encontra-se em andamento, existem outros objetivos que ainda serão desenvolvidos mais adiante. Com abordagem qualitativa, natureza aplicada e procedimentos de estudo de caso, a pesquisa completa desenvolve-se em sete etapas procedimentais, iniciando pela investigação e seleção de um padrão de metadados para representação de OA, a análise de repositórios existentes, a implementação de um ontologia operacional referente ao padrão de metadados escolhido, a integração interoperável em esquemas de metadados externos e posterior mapeamento para dados ligados com a sua publicação na Web semântica e, por fim, a realização de consultas para dados ligados de forma a demonstrar a recuperação de informação de alguns OA como prova de conceito da pesquisa. A primeira etapa já está concluída com a seleção do padrão LOM IEEE para representação de OA, e a segunda etapa parcialmente concluída. Espera-se que esta pesquisa, que encontra-se em andamento, contribua com a organização e representação de OA tornando-os mais encontráveis e acessíveis, e assim, mais propensos ao reuso.

Palavras-chave: Objetos de aprendizagem; Organização do conhecimento; Padrão LOM-IEEE; Dados Ligados; Web semântica.

Abstract
Due to the increase in courses in Distance Education, there has been an increase in the supply of Learning Objects (LO) that are normally available in repositories without interoperability with other bases, which generates redundancy of content and unnecessarily increases the cost of creation and maintenance , going against its main feature, which is reusability. In this context, this work proposes the investigation and selection of a metadata standard to represent LO, and also the investigation and analysis of the peculiarities of some LO available in selected repositories, making a survey of the necessary elements for its representation in the chosen metadata standard. As this research is in progress, there are other objectives that will still be developed later. With a qualitative approach, applied nature and case study procedures, the complete research is developed in seven procedural steps, starting with the investigation and selection of a metadata standard for representing LO, the analysis of existing repositories, the implementation of an ontology operational reference to the chosen metadata standard, interoperable integration into external metadata schemes and subsequent mapping to linked data with their publication on the semantic Web and, finally, performing queries for linked data in order to demonstrate the retrieval of information from some LO as proof of research concept. The first step is already completed with the selection of the IEEE LOM standard for LO representation, and the second step is partially completed. It is expected that this research, which is in progress, will contribute to the organization and representation of LO, making them more findable and accessible, and thus, more prone to reuse.

Keywords: Learning objects. Knowledge organization. LOM-IEEE standard. Linked Data. Semantic Web.

1 Introdução

Francisco e Braga (2021) destacam que os objetos de aprendizagem (OA) na EaD tornam o processo de aprendizagem mais prazeroso, desafiador e dinâmico, aguçando sua curiosidade e incentivando o estudante a buscar novos saberes além de promover sua interação e familiarização com os conteúdos

estudados. Behar (2009), destaca os OA como “recurso viável para enriquecer o espaço pedagógico” onde a sua utilização remete a um novo tipo de aprendizagem onde o professor deixa de ser apenas um transmissor de informações para ser um mediador da aprendizagem (BEHAR, 2009, p. 66). De acordo com Silva (2011), os OA ajudam na compreensão do todo, no que diz respeito à matéria de um curso, “atomizando” o conteúdo, ou seja, decompondo-o em partes menores de forma a ilustrar para o aprendiz como cada elemento contribui para o funcionamento do todo. “É como um relojoeiro que desmonta um relógio para um visitante, explicando as propriedades e funções de cada peça, individualmente, e o funcionamento do relógio como um todo” (SILVA, 2011, p. 17). Wiley (2003) usa metáfora dos blocos de montar tipo Lego para ilustrar o conceito de usabilidade dos OA popularizando o termo cunhado por Wayne Hodgins em 1994, porém, diferentemente deste, ressalta que nem todos os blocos combinam entre si, ou seja, nem todos os OA podem ser combinados com qualquer outro OA, tal como ocorre com esses blocos e propõe que os OA são mais como Átomos. Essas metáforas ressaltam a característica fundamental inerente aos OA que é a de organizar o conteúdo educacional em pequenos segmentos combinados entre si, formando resultados gradativamente mais complexos, de forma que possam ser reutilizados.

As Tecnologias Digitais de Informação e Comunicação (TDICs) têm contribuído na organização e catalogação dos OA com foco na melhoria do seu acesso e o reuso. Neste processo surgem os padrões de metadados que existem para categorizar, indexar, recuperar, reutilizar e combinar diferentes OA, que formam um conjunto mínimo de atributos de forma a permitir que estes objetos sejam caracterizados, localizados e avaliados. Os repositórios onde são armazenados os OA e disponibilizados em seus diversos formatos (textos, vídeos, áudios, imagens etc.), onde o metadado de um OA descreve as características mais relevantes utilizadas na sua catalogação nestes repositórios, permitindo assim, a sua recuperação por meio de sistemas de busca na Web ou mesmo num Ambiente Virtual de Aprendizagem. Contudo, para a cooperação entre computadores e pessoas ocorra, é importante a utilização dos princípios do *linked data*, ou dados ligados, que de acordo com Nhacuongue, Rozsa e Dutra (2018, p. 24) “é um conjunto de dados publicados na Web, em formatos que sejam legíveis por máquinas, com estruturas semânticas bem definidas, que estejam ligados a outros dados externos" permitindo o compartilhamento, reutilização e conexão com outros recursos e usuários. Processo que torna-se importante devido à expansão de cursos na educação a Distância (EaD) que tem aumentado a oferta e demanda por Objetos de Aprendizagem (OA) que normalmente estão disponíveis em repositórios sem interoperabilidade com outras bases, o que gera redundância de conteúdo e eleva desnecessariamente o custo de criação e manutenção, indo contrário à sua principal característica, que é a reusabilidade.

2 Objetivos

Os objetivos da pesquisa relatada no presente artigo são: [1] Investigar e selecionar um padrão de metadados para representar OA; e [2] Investigar e analisar as peculiaridades de alguns OA disponíveis em repositórios selecionados fazendo um levantamento dos elementos necessários para a sua representação no padrão de metadados escolhido.

Como esta pesquisa encontra-se em andamento, existem outros objetivos que ainda serão desenvolvidos mais adiante. Para contextualizar, segue a sua descrição, uma vez que alguns dos elementos metodológicos desses objetivos são apresentados na seção de procedimentos metodológicos: [3] Implementar o padrão de metadados escolhido em uma linguagem de representação ontológica na Web semântica; [4] Investigar elementos de esquemas de metadados externos e instâncias de outras bases de dados que possam contribuir para o aumento do nível de interoperabilidade da camada estrutural e camada semântica no padrão escolhido; [5] Mapear a implementação feita em [3], incorporando os elementos investigados em [4], para dados ligados interoperáveis na Web semântica; [6] Publicar o mapeamento de [5] como dados

ligados RDF na Web semântica; [7] Realizar buscas escritas em uma linguagem de consulta para dados ligados, como prova de conceito, a fim de validar a recuperação de informação dos OA publicados.

3 Procedimentos Metodológicos

Com abordagem qualitativa e natureza aplicada, a presente pesquisa utiliza-se de estudo de caso sobre dados de repositórios de OA que trabalham com Recursos Educacionais Abertos (REA).

A validação final acontecerá por meio de uma prova de conceito que, segundo Kendig (2016), refere-se a qualquer ideia que possa ser aplicada a uma classe de fenômenos, onde existe uma conexão causal hipotética na estrutura proposta, onde a função sugerida ou a abordagem metodológica adotada na pesquisa são obtidas em pelo menos um caso real (caso de teste) e acredita-se que os resultados obtidos com a experimentação num pequeno recorte, podem ser replicados em outras instâncias.

3.1 ETAPAS PROCEDIMENTAIS DO DESENVOLVIMENTO

Considerando a pesquisa como um todo, inclusive a parte que ainda está em andamento, existem sete etapas, sintonizadas com os objetivos. A primeira etapa já foi concluída. A segunda etapa foi parcialmente concluída, e as demais ainda encontram-se em desenvolvimento. Os próximos parágrafos apresentam as sete etapas, inclusive os processos e as ferramentas escolhidas até o presente momento e que subsidiarão o seu desenvolvimento.

Etapa 1: Investigação e seleção de um padrão de metadados para representar OA com melhor aderência aos OA, levando em consideração seus tipos, características e ciclo de vida.

Etapa 2: Investigação de repositórios de OA. Levantamento de repositórios de OA com base no tipo, característica e ciclo de vida de forma a investigar e analisar as informações necessárias para sua representação de acordo com o padrão escolhido na etapa 1.

Etapa 3: Implementação da ontologia operacional no domínio dos OAs. Implementação do padrão escolhido enquanto uma ontologia operacional que, segundo Falbo (2014), é uma uma das etapas da metodologia SABiO de desenvolvimento de ontologias de domínio, e tem foco na implementação da ontologia em uma linguagem específica. Pretende-se usar a linguagem RDF Turtle[1] baseada principalmente em OWL[2] e SKOS[3], e com o apoio do software Vocbench[4].

Etapa 4: Investigação e seleção de esquemas e padrões de metadados que possam contribuir para o aumento do nível de interoperabilidade da camada estrutural e camada semântica da ontologia operacional. Serão consultados portais de busca de vocabulários tais como o LOV[5], Prefix.cc[6] e o DCC[7]. A incorporação dos padrões selecionados na ontologia operacional, com apoio do software Vocbech.

Etapa 5: Mapeamento dos dados de alguns OA e ontologia operacional para uma base de dados ligados RDF interoperáveis na Web semântica.

Etapa 6: Publicação dos dados ligados que foram mapeados na Web.

---

[1] RDF Turtle é uma linguagem de marcação para representação de dados ligados RDF. Disponível em: https://www.w3.org/TR/turtle/.

[2] OWL é uma linguagem de modelagem usada para criar e/ou enriquecer ontologias. Disponível em: https://www.w3.org/OWL/.

[3] SKOS é um vocabulário recomendado pela W3C, projetado para representação de tesauros, esquemas de classificação, taxonomias, sistemas de controle de autoridade entre outros. Disponível em: https://www.w3.org/TR/skos-reference/.

[4] Vocbench é uma plataforma de desenvolvimento colaborativa baseada na web, multilíngue, para gerenciar ontologias OWL, tesauros SKOS(/XL), léxicos Ontolex-lemon e conjuntos de dados RDF genéricos. Disponível em: http://vocbench.uniroma2.it/.

[5] LOV (Linked Open Vocabularies) é um repositório de ontologias e portal de busca global para termos e vocabulários. Disponível em: https://lov.linkddata.es/dataset/lov/.

[6] Prefix.cc é um portal de busca de prefixos de namespaces. Disponível em: http://prefix.cc/.

[7] CDD (Digital Curation Centre) possui coletânea de esquemas de metadados organizados por disciplina. Disponível em: https://www.dcc.ac.uk/guidance/standards/metadata.

Etapa 7: Recuperação de informação por meio de escrita de buscas em uma linguagem de consulta para dados ligados a fim de testar os elementos desenvolvidos sobre os OA publicados.

4 Resultados

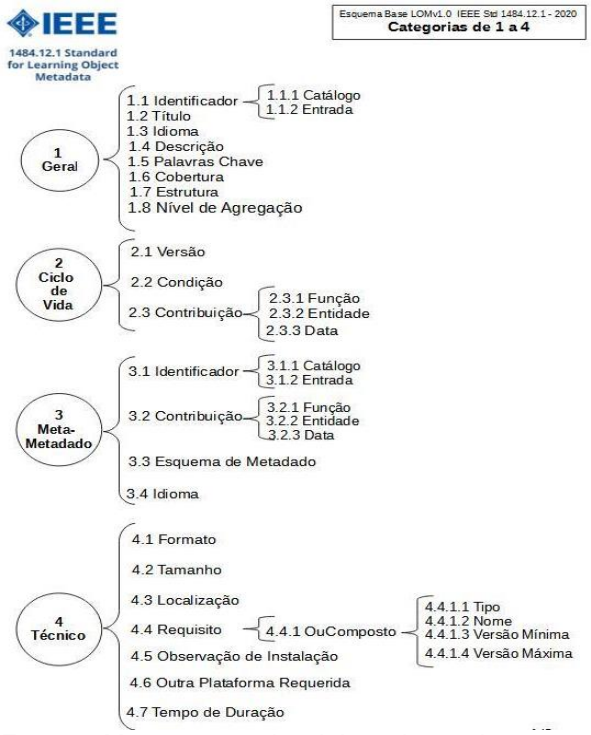
Por tratar-se de uma pesquisa em andamento, apresentar-se-á nesta seção os resultados relacionados aos padrões, processos e ferramentas escolhidos até o momento para o desenvolvimento de cada etapa de pesquisa aqui descritas.

Na etapa 1, ao investigar padrões de metadados que pudessem melhor representar os OA, primeiro buscou-se na literatura os conceitos relacionados à "Metadados", cujo uso é uma prática antiga da Biblioteconomia, fruto do desenvolvimento histórico de regras de catalogação que inclui padrões de conteúdo como AACR2, o seu sucessor, o RDA, e o caminho percorrido até a chegada do LRM, passando pelos modelos conceituais da família FR (FRBR, FRAD e FRSAD) e padrões de estrutura como os ISBD, o formato MARC Bibliográfico e BibFrame. Chegando-se aos mais atuais voltados para as novas Tecnologias Digitais de Informação e Comunicação - TDICs como DublinCore, linguagens eXtensible Markup Language - XML, Resource Description Framework - RDF e Ontology Web Language - OWL para representações de características de objetos digitais, especialmente na Web. (FOULONNEAU; RILEY, 2008; ZAFALON, 2012; MACHADO; ZAFALON, 2020).

Neste processo de pesquisa, chegou-se ao Learning Object Metadata - LOM (IEEE, 2020) um padrão internacional que é regido pela norma IEEE Std 1484.12.1™, 2020, cuja primeira versão foi aprovada em 12 junho de 2002 e teve sua última atualização em 24 de setembro de 2020. Trata-se de "um esquema de dados conceitual que define a estrutura de uma instância de metadados para um OA" (IEEE, 2020, tradução nossa). Ele traz em seu escopo a pretensão de que o padrão seja referenciado por outros padrões, onde propõe um esquema de dados conceitual que permite a diversidade linguística de modo que uma instância de um metadado de um OA possa ser utilizada por um sistema para gerenciar, localizar, avaliar ou trocar OA.

Há outros padrões para representação de objetos multimídias como o da W3C (World Wide Web Consortium) e da ISO/IEC (International Organization for Standardization/ International Electrotechnical Commission), que buscam soluções inteligentes para descrição de conteúdos multimídia processáveis por máquina e baseados em semântica como é o caso do padrão de metadados MPEG-7 ISO/IEC, comumente usado para descrição de conteúdo multimídia em rede.

Figura 1 – Esquema base do LOM/IEEE-2020 - categorias de 1 a 4



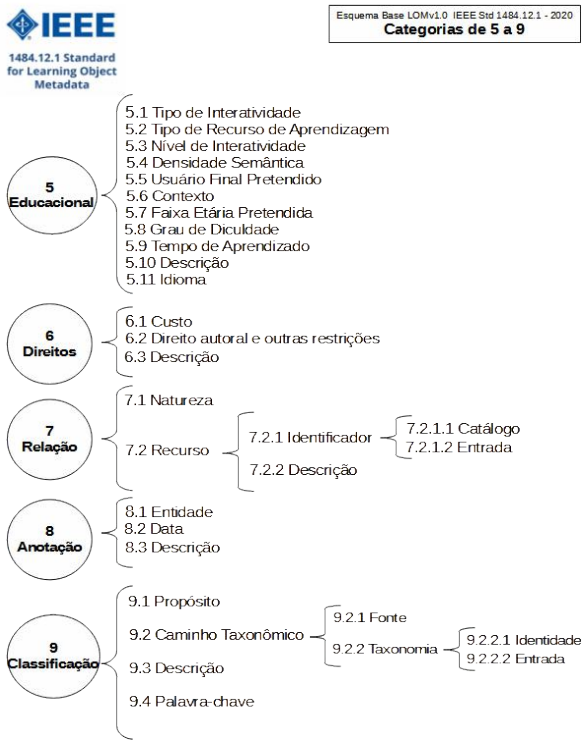
Fonte: Mapa mental elaborado pelo autor, adaptado de IEEE (2020)

Apesar da existência desses, optou-se pelo LOM por: [1] Facilitar a busca, avaliação, aquisição e uso de objetos de aprendizagem, por exemplo, por alunos, instrutores ou processos de software automatizados; [2] Facilitar o compartilhamento e troca de objetos de aprendizagem, permitindo o desenvolvimento de catálogos e inventários, levando em consideração a diversidade de contextos culturais e linguísticos em que os objetos de aprendizagem e seus metadados serão

explorados (IEEE, 2020, p.11, tradução nossa).

A ideia é que o LOM ajude a garantir que as ligações de Metadados de OA levem a um alto grau de interoperabilidade semântica e, como resultado, as transformações entre essas ligações sejam diretas. O esquema Base do LOM/IEEE-2020 consiste em nove categorias que descrevem os elementos de dados de um OA, de acordo com as Figuras 1 e 2.

Figura 2 – Esquema base do LOM/IEEE-2020 - categorias de 5 a 9



Fonte: Mapa mental elaborado pelo autor, adaptado de IEEE (2020)

Na etapa 2 (investigação de repositórios de objetos de aprendizagem), foram escolhidos os Repositório de Recursos Educacionais Abertos para Educação Profissional e Tecnológica - PROEDU, Banco Internacional de Objetos de Aprendizagem - BIOE e o ARCA – Repositório Institucional da Fiocruz por serem repositórios de acesso aberto. Porém, esta etapa ainda não foi concluída, pois ainda se está estudando os tipos de OA disponíveis nestas plataformas.

Na etapa 3 (implementação da ontologia operacional) após alguns estudos optou-se inicialmente pelo uso da plataforma VocBench para operacionalizar a criação da ontologia escrita em linguagem com suporte a RDF por ser uma plataforma de fácil aprendizado, com suporte a SKOS e que atende às necessidades da Web semântica e dados ligados (VOCBENCH, 2022).

A etapa 4 (investigação e seleção de esquemas e padrões de metadados) ainda está sendo desenvolvida.

Com vistas à execução da etapa 5 (mapeamento para dados ligados) foi escolhido o software GraphDB, por ser um banco de dados orientado a grafo e compatível com os padrões W3C. É um banco de dados da família RDF que são altamente eficientes, robustos e escaláveis. É compatível com vários padrões, como a especificação do protocolo W3C SPARQL e suporta todos os formatos de serialização RDF (GRAPHDB, 2022).

Na etapa 6 (publicação na Web semântica), pretende-se utilizar a Wikidata como plataforma para publicação, por ter um caráter colaborativo e agregar dados estruturados em relação a diversos elementos em diversas línguas. Além disso ela não requer qualquer conhecimento sobre linguagens de marcação, esquemas de dados, notação de objetos, ou outras sintaxes especiais, em vez disso, os dados são adicionados e editados diretamente em sua base por meio de formulários de entrada fáceis de usar (WIKIDATA, 2022).

Para a etapa 7 (recuperação de informação), pretende-se usar como linguagem de consulta em dados ligados a SPARQL, pois trata-se de uma linguagem padronizada para a consulta de grafos RDF, padrão desenvolvido pelo RDF Data Access Working Group do W3C (W3C, 2022). Contudo, essa linguagem não é apropriada para o usuário final.

5 Considerações Finais

Foram discutidas as razões para a seleção e escolha do padrão LOM IEEE para representação de OA. Mostrou-se a investigação de repositórios de objetos de aprendizagem que está sendo realizada com vistas a sua representação no padrão LOM IEEE. Também foram apresentadas as ferramentas selecionadas para desenvolvimento das etapas metodológicas.

Espera-se que a presente pesquisa, ainda em andamento, contribua com um direcionamento sobre organização e

representação de repositórios de OA com vistas a atender aos princípios FAIR[8]. Pois, "FAIRificar" estes recursos fará com que se tornem acessíveis, auxiliando os vários agentes envolvidos no processo ensino-aprendizagem (professores, alunos, designers instrucionais etc.) trazendo também, maior visibilidade ao trabalho riquíssimo desenvolvido nos ambientes virtuais de aprendizagem.

Principalmente, no princípio da reutilização/reusabilidade, apontada na literatura como a característica mais importante, assim, propor o uso do padrão LOM-IEEE no estabelecimento de um direcionamento dos metadados para catalogação de OA, faz-se com vistas a derrubar as barreiras que impedem a recuperação dos mesmos de forma integrada a outras bases. Além disso, publicá-los na Wikidata como dados ligados é como tirá-los das "caixinhas" dos repositórios e AVAs[9] e colocá-los de forma padronizada e interoperável na Web, ficando sintonizados ao mundo informacional altamente conectado, trazendo economia de recursos de mão de obra, tecnologia, tempo, e expandindo os horizontes para tornar a EaD mais efetiva pelo acesso e reuso facilitado de OA.

Observa-se ainda que ao término da pesquisa completa com as sete etapas, o uso da recuperação de informação ainda carece de uma interface ou camada de software apropriada, pois a linguagem proposta, SPARQL, é de difícil domínio por um usuário final.

[8] Acrônimo para Findable, Accessible, Interoperable e Reusable, que estão presentes nas discussões e práticas contemporâneas da Ciência da Informação desde o início de 2014, fruto de uma conferência internacional denominada "Jointly Designing a data FAIRPORT" (Jointly Designing a data FAIRPORT, 2014)

[9] Ambientes Virtuais de Aprendizagem, constituídos por uma infraestrutura tecnológica (interface gráfica, comunicação síncrona/assíncrona e outras funcionalidades) e por todas as relações (afetivas, cognitivas, simbólicas, entre outras) estabelecidas pelos participantes de cursos a distância. (BEHAR, 2009 p.202)

Referências

BEHAR, Patrícia Alejandra (Orgs.). Modelos pedagógicos em educação a distância. Porto Alegre: Artimed, 2009.

FALBO, Ricardo de Almeida. SABiO: Systematic approach for building ontologies. Em: 2014, Rio de Janeiro, RJ. Anais [...]. Em: 1st Joint Workshop ONTO.COM/ODISE on Ontologies in Conceptual Modeling and Information Systems Engineering co-located with 8th International Conference on Formal Ontology in Information Systems. Rio de Janeiro, RJ: CEUR Workshop Proceedings, 2014. p. 14. Disponível em: http://ceur-ws.org/Vol-1301/ontocomodise2014_2.pdf. Acesso em: 20 jul. 2022.

FOULONNEAU, M.; RILEY, J. Metadata for digital resources: implementation, systems design and interoperability. Oxford: Chandos, 2008.

FRANCISCO, Richard Fonseca; BRAGA, Antonio Celso de Oliveira. Objetos de aprendizagem: instrumentos para a avaliação formativa em educação a distância. Revista Paidéi@. Unimes Virtual. Volume 13 - Número 23. Janeiro – 2021. Disponível em: https://periodicos.unimesvirtual.com.br/index.php/paideia/index. Acesso em: 24 mar. 2022.

GRAPHDB. About GraphDB. Disponível em: https://graphdb.ontotext.com/documentation/10.0/about-graphdb.html. Acesso em: 20 jul. 2022.

IEEE - INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS. EEE1484.12.1-2020 - IEEE Standard for Learning Object Metadata - LOM. 2020. Disponível em <https://ieeexplore.ieee.org/servlet/opac?punumber=9262116>. Acesso em: 20 jul. 2022.

IMS. Instructional Management Systems. Disponível em: <https://www.1edtech.org/> Acesso em: 20 jun. 2022.

Jointly Designing a Data FAIRPORT, 2014, Workshop: 13 - 16 January 2014, Leiden, the Netherlands. Disponível em: https://www.lorentzcenter.nl/jointly-designing-a-data-fairport.html. Acesso em: 17 Mar. 2022.

KENDIG, C. E O. What is proof of concept research and how does it generate epistemic and ethical categories for future scientific practice? Sci Eng Ethics 22, 735–753 (2016). Disponível em: <https://link.springer.com/article/10.1007/s11948-015-9654-0>. Acesso em: 18 jun. 2022.

MACHADO, Raildo de Souza; I. ZAFALON, Zaira Regina. Catalogação: dos princípios e teorias ao RDA e IFLA LRM. João Pessoa: Editora UFPB, 2020.

NHACUONGUE, Januário Albino; ROZSA, Vitor; DUTRA, Moisés Lima. Linked Data e Ciência da Informação: diretrizes para a publicação de datasets institucionais abertos. Biblios, nº 73 (outubro): 20–34. 2018. Disponível em: <http://www.scielo.org.pe/scielo.php?script=sci_abstract&pid=S1562-47302018000400002&lng=es&nrm=iso&tlng=es>. Acesso em: 24 mar. 2022.

SILVA, Robson Santos da. Objetos de aprendizagem para a educação a distância. São Paulo: Novatec, 2011.

VOCBENCH. VocBench. Disponível em: http://vocbench.uniroma2.it/. Acesso em: 20 jul. 2022.

W3C. SPARQL Query Language for RDF. 2008. Disponível em: <https://www.w3.org/TR/rdf-sparql-query/>. Acesso em: 20 jul. 2022.

WIKIDATA. Welcome to Wikidata. Disponível em: https://www.wikidata.org/. Acesso em: 20 jul. 2022.

WILEY, D. A. The Post-LEGO Learning Object. 2003. Disponível em: <http://davidwiley.org/docs/post-lego.pdf>. Acesso em: 22 jun. 2022.

ZAFALON, Z. R. Scan for MARC: princípios sintáticos e semânticos de registros bibliográficos aplicados à conversão de dados analógicos para o Formato MARC21 Bibliográfico. 2012. Tese (Doutorado) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2012.

PATENTES COMO FONTE DE DADOS PARA ANÁLISE SOBRE A PRODUÇÃO TÉCNICA

PATENTS AS A SOURCE OF DATA FOR ANALYSIS ON TECHNICAL PRODUCTION

Raulivan Rodrigo da Silva[1], Thiago Magela Rodrigues Dias[2]
(1) CEFET/MG, R. Álvares de Azevedo, 400 - Bela Vista, Divinópolis - MG, 35503-822, raulivan@cefetmg.br
(2) CEFET/MG, R. Álvares de Azevedo, 400 - Bela Vista, Divinópolis - MG, 35503-822, thiagomagela@cefetmg.br

Resumo
Este trabalho busca contribuir com a compreensão do cenário tecnológico nacional, utilizando dados provenientes de patentes como objeto de análise. Objetivando contribuir com o projeto BRCRIS do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), propõem-se uma estratégia para coleta de patentes depositadas no Instituto Nacional da Propriedade Industrial (INPI), além de identificar proponentes de patentes na base curricular da Plataforma Lattes. A metodologia proposta, consiste em coletar as patentes depositadas no Brasil utilizando um repositório internacional de publicação de dados de patentes, a Espacenet, que contém dados de patentes de mais de 70 países. Posteriormente, por meio do conjunto ferramental proposto, obtêm-se os currículos da Plataforma Lattes que possuem dados de patentes informada, viabilizando certificar com dados da Espacenet um conjunto com 31.816 registros com informações de patentes em 16.445 currículos da Plataforma Lattes.

Palavras-chave: Patente; BRCRIS; Espacenet; Plataforma Lattes; Coleta de dados.

Abstract
This work seeks to contribute to the understanding of the national technological scenario, using data from patents as an object of analysis. Aiming to contribute to the BRCRIS project of the Brazilian Institute of Information in Science and Technology (IBICT), a strategy for collecting patents deposited at the National Institute of Industrial Property (INPI) is proposed, in addition to identifying patent applicants in the curricular base of the Platform Lattes. The proposed methodology consists of collecting patents deposited in Brazil using an international repository for publishing patent data, Espacenet, which contains patent data from more than 70 countries. Subsequently, through the proposed toolkit, the Lattes Platform curricula that have informed patent data are obtained, making it possible to certify with Espacenet data a set with 31,816 records with patent information in 16,445 Lattes Platform curricula.

Keywords: Patent; BRCRIS; Espacenet; Lattes Platform; Data collect.

1 Introdução

O século XXI tem sido solo fértil para a criação de estruturas tecnológicas e, mais do que nunca, a rapidez na evolução destas tecnologias tem sido visível. Com o advento das redes de compartilhamento de informações por meio da internet, é possível deparar com um expressivo volume de dados advindos produções científicas.

Mediante ao exposto, é importante mensurar toda essas informações produzidas para acompanhar o progresso científico e tecnológico, bem como contribuir para sua evolução. Os Estudos Métricos da Informação são uma das áreas de interesse da Ciência da Informação e tem como foco a identificação e avaliação da informação, seu alcance, influência e impacto, de acordo com Cabrini e Gracio (2011), os "Estudos Métricos" são constituídos por um conjunto de estudos relacionados à avaliação da informação produzida.

De acordo com o foco de interesse, da natureza da informação e do objeto de análise, os ramos dos estudos métricos podem ser classificados como Bibliométricos, Informétricos ou Infométricos, Cientométricos, Cibermétricos, Webométricos, Patentométricos e Arquivométricos (CURTY; DELBIANCO, 2020).

No contexto da produção técnica, documentos de patentes se apresentam como uma rica fonte de informação tecnológica. A compreensão do estado da técnica da arte por meio de documentos de patentes, consequentemente apresenta um cenário mais assertivo a respeito de

tendências tecnológicas, setores promissores, bem como, a possibilidade de novas tecnologias (NASCIMENTO; SPEZIALI, 2020). Os estudos e análise de documentos de patentes permitem identificar o conhecimento científico e convertê-lo em conhecimento tecnológico.

Conforme apontam Nascimento e Speziali (2020) o mapeamento de tecnologias utilizando informações contidas em documentos de patentes, é pouco explorado no Brasil, salientando que, por fazer parte de áreas estratégicas de muitas empresas, relatórios de mapeamentos tecnológicos mais completos não são disponibilizados de forma gratuita para consulta. Calzolaio et al. (2018) e Tanaka e Inui (2016) concordam que os dados de patentes contêm informações valiosas para análises técnicas, entretanto, esta é uma área pouco explorada também pelas universidades. A prospecção tecnológica é uma área de estudo recente principalmente no Brasil, que possui uma literatura limitada sobre o tema. Serão aceitos estudos concluídos e em andamento, desde que tenha resultados e conclusões, mesmo que parciais.

Nessa conjuntura, o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), desde 2014 está desenvolvendo o *Brazilian Current Research Information System* (BrCRIS), um ecossistema de sistemas, com objetivo de reunir dados de diversas fontes tais como dados de Projetos de pesquisa, financiamento, pesquisadores, infraestrutura de pesquisa, instituições de pesquisa e seus outputs em C&T, constituindo um conjunto de dados para criação de sistemas de recomendações em dados abertos.

2 Objetivos

O objetivo principal deste trabalho é apresentar uma estratégia de coleta de patentes depositadas no Instituto Nacional da Propriedade Industrial[1] (INPI) no repositório internacional Espacenet[2]. Ensejando contribuir com o desenvolvimento do BRCRIS, certificando os dados de patentes, cujo seus respectivos proponentes, informaram registros de patentes em seus currículos da Plataforma Lattes.

3 Procedimentos Metodológicos

A metodologia foi dividida em três partes distintas, a saber, (1) Coleta de dados de Patentes; (2) Coleta dos dados da Plataforma Lattes; e por fim (3) Tratamento e Seleção dos dados.

A primeira parte consiste na coleta de patentes brasileiras disponíveis na Espacenet, ou seja, patentes que foram depositadas no INPI até o ano de 2021 e disponibilizadas para consulta na Espacenet. A coleta dos dados foi realizada no primeiro semestre do ano de 2022. Para otimizar o processo de coleta foi desenvolvido um script utilizando a linguagem de programação Python, para acessar a OPS (*Open Patent Services*), um serviço web que a Espacenet disponibiliza para fornecer acesso à base de dados de patentes. Ressaltando que para consumir tais serviços é necessário realizar um cadastro no site da Espacenet para obter as credenciais de acesso (Espacenet, 2020). Fazendo uso do serviço de busca, é possível pesquisar as patentes depositadas em um determinado período, por exemplo, para consultar patentes brasileiras publicadas no período de janeiro de 2021 à dezembro de 2021, se faz necessário construir a consulta da seguinte forma: *search?q=pn=BR and pd="20210101 20210131",* em que *search* é serviço de consulta de patente. O parâmetro *q* que recebe os critérios da consulta, já no parâmetro *pn* é informado o código do país no número da publicação e por fim, utilizando o conector "*and*" é informado o parâmetro *pd* em que é informado o período de publicação da patente, destacando a data inicial e final que devem ser informadas no formato AAAAMMDD (ano, mês, dia).

Após obter a lista de patentes brasileiras disponíveis na Espacenet, se fez necessário realizar o obter dos dados das patentes, usando o serviço "*publication/epodoc/{número-de-depósito }/biblio*", onde foi informado o número de depósito de cada patente no formato EPODOC, este último é uma regra de formatação aplicado no número de identificação da patente.

Todas as patentes foram armazenadas em arquivos no formato .json (*JavaScript*

[1] <https://www.gov.br/inpi/pt-br>

[2] <https://worldwide.espacenet.com>

Object Notation), em um diretório denominado "PATENTESBR". O nome de cada arquivos é formado pelo número da patente o qual armazena os dados, por exemplo “BR0107786A.json”, com base nesta informação, foram criados sub diretórios para armazenar as patentes, cujo o nome do subdiretório é composto pelos 4 primeiros caracteres do nome do arquivo, de acordo com o exemplo citado, o nome do diretório a qual ele pertence é “BR01”.

Dando sequência, a segunda etapa foi coletar currículos registrados na Plataforma Lattes que possuem informações de patentes, como o número do pedido de depósito ou o título da patente. O processo de coleta e seleção dos dados curriculares da Plataforma Lattes foi realizado por meio do framework *LattesDataXplorer* (DIAS, 2016). O framework possui um conjunto de técnicas e métodos responsáveis por coletar, selecionar, tratar e analisar os dados da Plataforma Lattes.

O extrator coleta os currículos e os armazena no formato XML em pastas, identificadas de 00 a 99. O nome da pasta selecionada para armazenar o arquivo e o nome do mesmo são definidas de acordo com seu número único de identificador de 16 dígitos, os dois primeiros números do identificador correspondem ao nome da pasta e os 14 dígitos restantes correspondem ao nome do arquivo salvo. A coleta dos currículos foi realizada no primeiro semestre de 2022.

Nos currículos da Plataforma Lattes o número de depósito de patente, bem como as demais informações, são inseridas pelo o próprio pesquisador, o que geralmente ocasiona uma falta de padrão no registro das informações. Portanto, se faz necessário implementar o terceiro passo que consiste em tratar os números de depósito informado nos currículos coletados para possibilitar a identificação dos mesmos na base coletada na Espacenet. Este procedimento consiste em realizar uma sequência de passos, a saber: o “passo-1” da consiste na remoção da formatação do número de depósito, removendo todos os caracteres especiais como ponto, vírgula, símbolos, dentre outros. Em continuidade no tratamento do número de depósito, no “passo-2” é realizada a remoção do último dígito que compõem o número. Feito isto, o “passo-3” consiste em uma busca aproximada pelo o número de depósito já tratado, esta busca aproximada consiste em pesquisar nos arquivos de patentes o número de depósito usando como critério de seleção: %[número tratado]%, em que as porcentagens representam qualquer caractere, ou seja, qualquer que seja o valor antes e/ou depois do número tratado será considerado como resultado satisfatório da busca. Caso tenha localizado alguma patente que atenda aos critérios de busca no “passo-3”, é realizado o “passo-4 ”, em que é verificado se o nome do pesquisador consta na lista de inventores da patente, para isso, usa-se o nome do pesquisador conforme foi informado em seu currículo, a exemplo “Raulivan Rodrigo da Silva”, posteriormente o nome é dividido por espaços formando uma lista, com base no exemplo dado, a lista ficaria com 4 elementos “[Raulivan, Rodrigo, da, Silva]”, feito isto, verifica se todos os elementos da lista correspondem aos nome dos inventores, respeitando a mesma grafia e ordem de ocorrência, ignorando outros possíveis termos existentes no nome dos inventores. Caso tenha encontrado, finaliza-se o processo considerando o número informado pelo pesquisado como válido, caso não tenha encontrado, finaliza-se o processo considerando o número informado pelo pesquisador como inválido. Ainda no “passo-3” existe um fluxo alternativo, em que, caso não tenha localizada uma patente é investigado se o número utilizado como critério de busca inicia com alguns dos prefixos “CI”,”DI”,”UM” ou “PI”, em caso negativo, finaliza o processo considerando o número informado inválido, mas em caso afirmativo, é feita a substituição do prefixo identificado por “BR”, voltando assim ao “passo-3” dando continuidade no fluxo já estabelecido.

Para automatizar o processo de validação, foi implementado um algoritmo utilizando a linguagem de programação Python.

4 Resultados

Os dados coletados da Espacenet totalizaram 858.622 registros de patentes, contendo 1.914.866 nomes de depositantes e 4.386.733 nomes de inventores, cerca de 25,6 Gb (Gigabyte) de dados. O quantitativo

de patentes coletadas na Espacenet corresponde aproximadamente a 90,83% do conjunto de patentes depositadas no INPI.

Para apresentar um panorama da evolução tecnológica nacional, foi realizada a análise dos depósitos anuais de patentes, apresentando dados entre 1972 a 2021, data de registro da patente mais antiga até o último ano de análise. A **Figura 1 - Apêndice A** apresenta a evolução temporal no número de depósito de patentes, ressaltando que seis patentes não contém a informação de ano de depósito.

Cada patente de acordo com sua natureza e finalidade recebe uma classificação de acordo com o sistema internacional de classificação de patentes, com base nestas classificações é possível compreender quais áreas do conhecimento têm gerado o maior número de patentes. Para compreender melhor este cenário foi compilado no gráfico, apresentado pela **Figura 2 - Apêndice A**, as classificações recebidas pelas patentes brasileiras consideradas neste estudo.

Dentro do contexto da Plataforma Lattes, os resultados foram obtidos mediante a análise dos registros de patentes contidos nos currículos cadastrados. Atualmente a Plataforma Lattes é composta por mais de 7.4 milhões de currículos, que abrange indivíduos nos diversos níveis de formação acadêmica, no entanto, somente 29.516 possuem informações de patentes registradas, menos de 1% de toda a base de dados curriculares. Os 29.516 currículos possuem juntos um total 72.256 registros com informações de patentes, contudo, não foram todos considerados, apenas 31.816 registros foram devidamente identificados na base de dados coletada na Espacenet, totalizando 16.445 currículos. O restante não foi possível identificar na base de dados da Espacenet aplicando as estratégias definidas neste estudo. De encontro com objetivo deste estudo, todo conjunto de dados de patentes dos 16.445 currículos foram incorporados ao projeto BRCRIS no segundo semestre de 2022.

5 Considerações Finais

Este estudo abordou a perspectiva da ciência da informação, expondo os dados provenientes de patentes como uma fonte confiável e ampla no que se refere ao desenvolvimento tecnológico nacional.

Neste contexto, mediante aos resultados obtidos, é possível concluir que todos os objetivos apresentados neste estudo foram alcançados. A estratégia de coleta de dados de patentes proposta, viabiliza manter a base de dados sempre atualizada, executando a mesma sempre atualizando o período de depósito desejado, permitindo obter os dados de patentes de anos subsequentes da mencionada neste estudo. Um fato relevante a ser considerado, que o processo de coleta não é um processo rápido de ser executado, a Espacenet impõe limites de coletas mensais, que quando são atingidos, é necessário parar a coleta e esperar a próxima semana, pois os limites são renovados a cada domingo. Outro ponto, que foi identificado durante a realização deste estudo, é que é recomendado executar a coleta fora do horário comercial, pois durante o horário comercial, quando se realiza muitas requisições à API da Espacenet, sua conta é bloqueada por algumas horas, voltando a funcionar corretamente.

Já considerando os dados da Plataforma Lattes, apenas cerca de 1% de todos os currículos da base de dados da Plataforma Lattes, possuem informações sobre o depósito de patentes, base composta por mais de 7.4 milhões de currículos. Do conjunto de registros de patentes recuperados dos currículos, nem todos puderam ser validados na Espacenet devido à inconsistência nos dados registrados, notabilizando a necessidade da existência de mecanismos de validação e certificação dos dados patentários informados pelos os proprietários dos currículos. Contudo, foi possível contribuir com o desenvolvimento do projeto BRCRIS, identificando proponentes de patentes na base de dados curriculares da Plataforma Lattes, consequentemente fornecendo os dados das patentes identificadas.

Referências

CABRINI, E. F. T. de O. M. C.; GRACIO. Indicadores bibliométricos em ciência da informação: análise dos pesquisadores mais produtivos no tema estudos métricos na base scopus. **Perspectivas em Ciência da**

Informação, v. 16, p. 16–28, out. 2011. Disponível em: <https://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362011000400003&

lng=pt&tlng=pt>. Acesso em: 16 set. de 2022.

CALZOLAIO, A. P. A. E. et al. Mapeamento dos registros de propriedade intelectual (patentes) na universidade federal do rio grande do sul. **Revista Brasileira de Gestão e Inovação**, v. 6, n. 1, p. 44–70, 2018. Disponível em: <http://www.ucs.br/etc/revistas/index.php/RBGI/article/view/5860>. Acesso em: 16 set. de 2022.

CURTY, N. R. R. G.; DELBIANCO. As diferentes métricas dos estudos métricos da informação: evolução epistemológica, inter-relações e representações. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 25, p. 01–21, 2020.

DIAS, T. M. R. **Um Estudo da Produção Científica Brasileira a partir de Dados da Plataforma Lattes**. 181 p. Tese (Doutorado em Modelagem Matemática e Computacional) - Centro Federal de Educação Tecnológica de Minas Gerais, set. 2016.

NASCIMENTO, M. G. Raphael da S.; SPEZIALI. Patentometria: a utilização de dados contidos em patentes como mecanismo de análise da predominância tecnológica dos nits. **IV Encontro Internacional de Gestão, Desenvolvimento e Inovação**, nov. 2020.

NASCIMENTO, M. G. Raphael da S.; SPEZIALI. Patentometria: a utilização de dados contidos em patentes como mecanismo de análise da predominância tecnológica dos nits. **IV Encontro Internacional de Gestão, Desenvolvimento e Inovação**, nov. 2020.

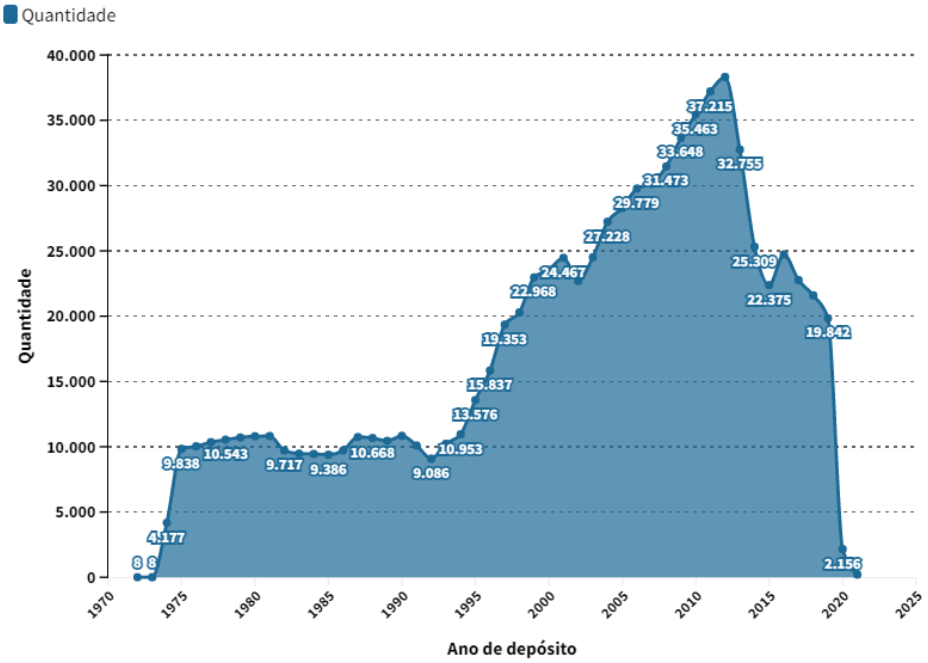
Espacenet. OPS. Open Patent Services RESTful Web Services. 1.3.16. ed. Disponível em: <https://www.epo.org/searching-for-patents/data/web-services/ops.html>. Acesso em: 13/01/2022.

PINTO, Adilson Luiz; SEGUNDO, Washington Luís Ribeiro de; QUONIAM, Luc; DIAS, Thiago Magela Rodrigues. Atas do V Congresso ISKO Espanha-Portugal: BRCRIS. In: **Organização do Conhecimento no Horizonte 2030: Desenvolvimento Sustentável e Saúde**. 2021. cap. The brazilian current research information system, p. 319-330. ISBN 978-989-566-137-4. Disponível em: <https://dialnet.unirioja.es/servlet/articulo?codigo=8411204>. Acesso em: 16 de set. 2022.

TANAKA, T. Y.; INUI. Preliminary study on why university researchers do not utilize patente information for their academic research in the field of science and engineering in japan. **Portland International Conference on Management of Engineering and Technology (PICMET)**, p. 1609–1618, 2016.

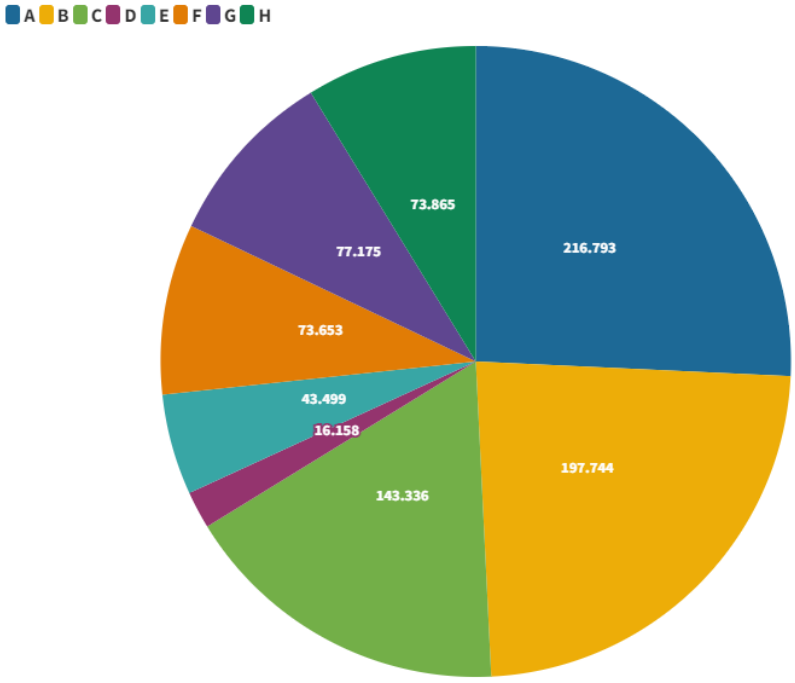
Apêndice A – Figuras

Figura 1 – Evolução temporal do depósito de patentes por ano



Dados da Pesquisa, 2022.

Figura 2 – Patentes por classificação



A - Necessidades humanas; B - Operações de processamento, transportes; C - Química; Metalurgia; D – Têxteis, Papel; E – Construções fixas; F – Engenharia mecânica, Iluminação, Aquecimento, Armas, Explosões; G – Física; H – Eletricidade.
Dados da Pesquisa, 2022

PLATAFORMA PARA AGREGAÇÃO E ANÁLISES DE DADOS TÉCNICOS-CIENTÍFICOS

PLATFORM FOR AGGREGATION AND ANALYSIS OF TECHNICAL SCIENTIFIC DATA

Thiago Magela Rodrigues Dias[1], Washington Luís R. de Carvalho Segundo[2], Tales Henrique José Moreira[2], Vivian dos Santos Silva[2], Adilson Luiz Pinto[3]

(1) CEFET/MG, R. Álvares de Azevedo, 400 - Bela Vista, Divinópolis - MG, 35503-822, thiagomagela@cefetmg.br

(2) IBICT, SAUS Quadra 5, Lote 6, Bloco H, Brasília - DF, 70070-912, washingtonsegundo@ibict.br, tales.info@gmail.com, viviansilva@ibict.br

(3) UFSC, R. Eng. Agronômico Andrei Cristian Ferreira s/n- Trindade, Florianópolis - SC, 88404-900, adilson.pinto@ufsc.br

Resumo

Nos últimos anos várias iniciativas que visavam a criação de sistemas que gerenciam a produção acadêmica de uma instituição, país ou área de conhecimento tem recebido atenção de diversas áreas. Tais sistemas são conhecidos pela sigla CRIS (Current Research Information Systems) e têm como objetivo agregar informações de bases de dados diversas com intuito de fornecer relatórios e dados consolidados para que pesquisadores possam analisar. Logo, este trabalho descreve a evolução do tratamento de dados que são fonte de informação para a plataforma BrCris no intuito de fornecer ferramentas tecnológicas visando munir a comunidade científica brasileira com dados consolidados da produção nacional.

Palavras-chave: BRCRIS; Plataforma Lattes; Tratamento de dados.

Abstract

In recent years, several initiatives aimed at creating systems that manage the academic production of an institution, country or area of knowledge have received attention from several areas. Such systems are known by the acronym CRIS (Current Research Information Systems) and aim to aggregate information from different databases in order to provide consolidated reports and data for researchers to analyze. Therefore, this work describes the evolution of the treatment of data that are a source of information for the BrCris platform in order to provide technological tools with the aim of providing the Brazilian scientific community with consolidated data from national production.

Keywords: BRCRIS; Lattes Platform; Data processing.

1 Introdução

A produção científica brasileira tem crescido expressivamente e, em perspectiva às especificidades de campos disciplinares distintos, heterogênea quanto à tipificação de sua produção tanto em termos quantitativos como qualitativos. E o resultado desta produção se materializa em forma de artigos em periódicos, teses e dissertações, além de produtos diversos como: softwares, patentes, obras e instalações artísticas, entrevistas e projetos cinematográficos.

Para o campo da Ciência da Informação, e em especial da Cientometria, quantificar essa produção é uma tarefa árdua pois a disponibilização de bases de dados abertas muitas vezes é restrita ou simplesmente inexistente em determinados contextos. Bases de dados proprietárias como a Scopus, Web of Science, Google Acadêmico e Microsoft Research Data permitem o acesso, mas este é sempre limitado ao número de registros que podem ser obtidos, contemplam poucos repositórios e periódicos nacionais e ainda existe o grave problema da opacidade dos algoritmos utilizados por estas plataformas que determinam o que é ou não relevante.

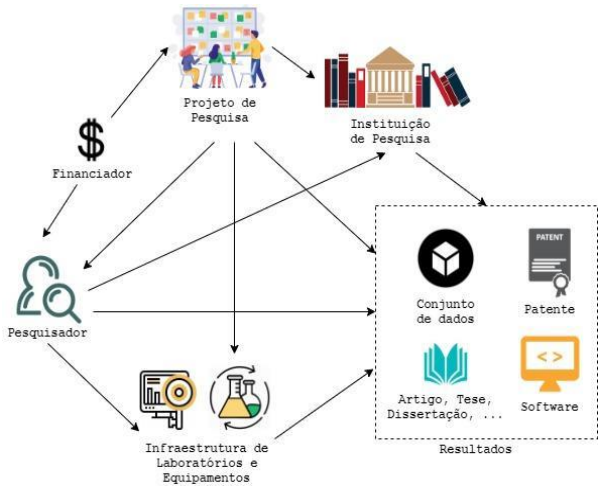
A partir desse cenário, começaram a surgir iniciativas que visavam a criação de sistemas que gerenciam a produção acadêmica de uma instituição, país ou área de conhecimento. Tais sistemas são conhecidos pela sigla CRIS (*Current Research Information Systems*) e têm como objetivo agregar informações de bases de dados diversas com intuito de fornecer relatórios e dados consolidados para que pesquisadores da área possam analisar

como se dá a produção em seus países ou áreas de conhecimento.

CRIS define um sistema de informação sobre todo o ecossistema do processo científico. São organizadas em um só lugar todas informações do ciclo da pesquisa Científica, desde o Fomento, passando pelos Projetos, Pesquisadores, Instituições de Pesquisa e Laboratórios, até os outputs de uma pesquisa científica, tais como artigos científicos, teses, dissertações, livros, capítulos de livro, patentes e conjuntos de dados científicos (SIVERTSEN, 2019).

Neste contexto, a idealização do Projeto do Sistema BrCris (PINTO et al., 2021), que é o CRIS no contexto da Ciência Brasileira, data de 2014, quando inspirado no modelo proposto por Portugal de um CRIS nacional (o PTCRIS - https://ptcris.pt), o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) iniciou uma sequência de estudos e parcerias interinstitucionais para a execução do Projeto. Em 2020, houve a implementação formal de um Projeto de Pesquisa para a construção do BrCris. O intuito é fornecer ferramentas tecnológicas visando munir a comunidade acadêmica brasileira com dados consolidados da produção científica nacional. Tomando como base outros projetos CRIS e padrões internacionais disponibilizados pelo OpenAire e COAR.

Figura 1 – Ecosistema da Pesquisa Científica.



Dados da Pesquisa, 2022.

Logo, o BrCris tem por objetivo estabelecer um modelo único de organização da informação científica de todo o ecossistema da pesquisa brasileira. Entre os agentes deste ecossistema estão os pesquisadores, os projetos, infraestruturas, laboratórios e instituições de pesquisa, os financiadores (KONG, et al., 2019), além dos resultados da pesquisa expressos principalmente por publicações científicas, teses, dissertações, conjuntos de dados científicos, software e patentes (ver Figura 1).

Diante disso, com a integração dos dados em um repositório de dados padronizado, o uso do dashboard em forma de visualização elucida alguns benefícios, como a redução de complexidade de dados, auxilia na percepção das propriedades existentes, ajuda na detecção de erros aparentes, consegue englobar a representação em pouco conteúdo, amplia a percepção cognitiva, dentre outros.

Outro aspecto que vale ser salientado na função do dashboard é a geração de índice e indicadores visando atribuir a mensuração de fenômenos, sejam de natureza social, econômica ou científico/tecnológica. No contexto deste trabalho, valoriza-se o foco em sua representação e facilidade de identificação dos cenários.

Tendo em vista o que foi exposto, este trabalho tem como objetivo apresentar o processo de desenvolvimento do projeto BrCris, que visa coletar, integrar e disponibilizar informações relativas ao universo de pesquisa científica no Brasil, catalogando e traçando relacionamentos entre pesquisadores, as organizações às quais pertencem, os projetos dos quais participam e como são financiados, e todos os produtos por ele gerados, como publicações, patentes e software.

2 Desenvolvimento

O BrCris concentra um amplo ecossistema de dados, de diversas fontes, como por exemplo, dados curriculares de indivíduos, sobre organizações, programas de pós-graduação, publicações, orientações, revistas científicas, dentre outros, sendo necessário todo um esforço para tratamento dos dados de interesse. Neste contexto, tendo em vista as diversas fontes de dados que irão compor o BrCris se faz necessário a transformação dos dados em formato padronizado, sendo necessário a

transformação baseada em um modelo que será importado para a plataforma *LA Referencia*.

Em se tratando do modelo de dados do BrCris, iniciou-se pela adoção de nove entidades de dados, seguindo padrões amplamente utilizados na comunidade científica internacional. São elas:

- Project: projetos de pesquisa executados, ou em execução;
- Service: revistas científicas, repositórios digitais, bibliotecas digitais e outras fontes de informação científica;
- Program: programas de pós-graduação brasileiros;
- Course: cursos nacionais, ou internacionais de pós-graduação stricto ou lato sensu;
- OrgUnit: instituições, faculdades, departamentos de pesquisa;
- Person: pesquisadores, assistentes de pesquisa e pessoas de apoio técnico à pesquisa;
- Patent: patentes como resultado da pesquisa;
- Dataset: conjuntos de dados de pesquisa coletados por pesquisadores e demais agentes no âmbito de um projeto ou pesquisa científica;
- Publication: artigos científicos, teses, dissertações, livros, capítulos de livro e relatórios científicos.

Figura 2 – Entidades do descritivo.

Field	Description	Diadorim	DOAJ
identifier.diadorim	id diadorim	id	
identifier.latindex	id LatinDex		
identifier.ulrichsweb	id ULrichsweb		
identifier.issn	id issn	dc.identifier.issn[pt_BR]	Journal EISSN (online version)
identifier.issnl	id issnl	dc.identifier.issn[pt_BR]	Journal ISSN (print version)
identifier.oai	id oai	dc.identifier.oai	
identifier.uri	id uri	dc.identifier.uri	Journal URL
identifier.brcris	hash gerado com título + publisher	dc.title[pt_BR] + dc.publisher.name[pt_BR]	Journal title + Publisher
identifier.eissn	(Será considerado como um issn)		Journal EISSN (online version)
identifier.opendoarId			
identifier.other			
compatibility			
acronym			
status	status da revista (active / inactive)	dc.relation.situation[pt_BR]	
accessType	tipo (direito) de acesso	dc.rights.access[pt_BR]	Permite acesso ao texto completo
ccLicence	Permissões	dc.rights.cc[pt_BR]	Journal license
rightsType		dc.rights.type[pt_BR]	
researchArea	área de pesquisa		

Dados da Pesquisa, 2022.

O modelo de dados é definido por um conjunto de entidades e relações, que por sua vez possuem identificadores e atributos pré-definidos. A utilização de um descritivo visa facilitar a identificação de atributos de cada entidade (Figura 2) e suas relações (Figura 3), possibilitando que com o auxílio de uma rotina desenvolvida especificamente para esta funcionalidade, o modelo possa incorporar todas as mudanças realizadas diretamente no modelo. Esta estratégia visa facilitar de forma significativa a incorporação de novos atributos e relações, sem a necessidade de alterações diretamente no modelo de dados.

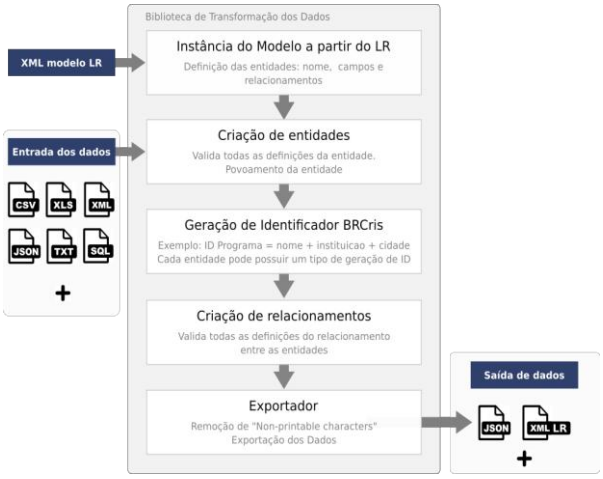
Figura 3 – Relações do Descritivo.

Name	From Entity	From Label	To Entity	To Label	Description
Affiliation	OrgUnit	hasMember	Person	isMemberOf	Is a relation between Person and OrgUnit.
IsUnitOf	OrgUnit	isUnitOf	OrgUnit	hasUnit	The unit related to an organization.
OrgUnitProgram	OrgUnit	hasProgram	Program	isProgramOf	The program related to an organization.
OrgUnitProject	OrgUnit	hasProject	Project	isProjectOf	Is a relation between Project and OrgUnit.
ThesisSponsorship	OrgUnit	sponsors	Publication	isSponsoredBy	Is a relation between a publication of type thesis and a sponsor OrgUnit.
CourseOrgUnit	OrgUnit	hasCourse	Course	isProvidedBy	The OrgUnit(s) that provided the Course.
Authorship	Publication	hasCreator	Person	isCreatorOf	The author of this content or rating.
Advisoring	Publication	hasAdvisor	Person	isAdvisorOf	The advisor of this content or rating.
CoAdvisoring	Publication	hasCoAdvisor	Person	isCoAdvisorOf	The coadvisor of this content or rating.
Referee	Publication	hasRefeee	Person	isRefereeOf	The advisor of this content or rating.
Publisher	Publication	hasPublisher	Senice	isPublishedIn	The editor of this content or rating.
PartOf	Publication	hasPartOf	Publication	isPartOf	The publisher of this content or rating.
ProgramThesis	Publication	hasPublication	Program	isProgramOf	Is a relation between two publications.
SeniceOrgUnit	Senice	hasOrgUnit	OrgUnit	isOrgUnitOf	The publication (thesis) related to an program.
Editoring	Senice	hasEditor	Person	isEditorOf	The Course that is associated with a thesis (publication)
CourseThesis	Course	isAssociatedTo	Publication	hasAssociationWith	Is a relation between Patent and Person.
Inventor	Patent	hasInventor	Person	isInventorOf	Is a relation related to an OrgUnit of Patent.
PatentOrgUnit	Patent	isPatent	OrgUnit	hasPatent	Is a relation between Project consortion and member.

Dados da Pesquisa, 2022.

Como pode ser observado, o descritivo de uma Entidade apresenta inicialmente o atributo definido no Modelo de Dados bem como sua respectiva descrição. Logo, para cada conjunto de dados que é fonte de informações para a Entidade, são descritos os atributos dos conjuntos de dados relacionados aos do modelo.

Figura 4 – Estrutura da biblioteca de tratamento dos dados.



Dados da Pesquisa, 2022.

Para o tratamento dos dados foi desenvolvida uma biblioteca na linguagem de programação Python (Figura 4), contendo uma estrutura de dados preparada para facilitar o processamento de dados originários de todas as fontes para o formato

exigido pela plataforma *LA Referencia*. Logo, a biblioteca desenvolvida é responsável por toda a transformação e exportação dos dados, utilizando como base o "Modelo de Dados" da plataforma LA Referencia, validando as entidades, campos e relacionamentos aceitos pelo modelo.

Além disso, a biblioteca desenvolvida também é responsável por gerar Identificadores BrCris, criados com o intuito de realizar uma pré-desambiguação dos dados, evitando entidades duplicadas na plataforma *LA Referencia*. A geração do Identificador BrCris é realizada de forma distinta para cada entidade, em geral realizando um hash de seus próprios campos de dados, como por exemplo nas entidades:

- OrgUnit: hash a partir da concatenação do nome da organização e nome da cidade de localização;
- Publication: hash a partir da concatenação do título, tipo e ano da publicação;
- Program: hash a partir da concatenação do nome do programa de pós-graduação, instituição e cidade de localização;
- Service: hash a partir da concatenação do país e do nome da editora de um periódico científico.

Por fim, o ferramental proposto possibilita realizar a exportação dos dados originais no formato XML (Extensible Markup Language), no padrão que será importado pela plataforma LA Referencia

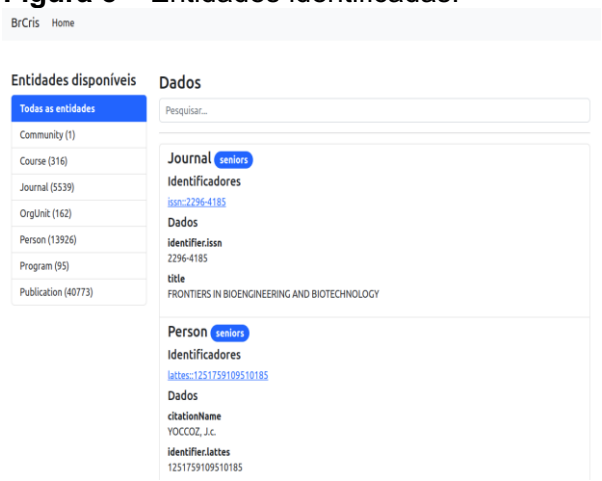
3 Resultados

Os resultados da execução do Projeto já incluem o desenvolvimento da arquitetura do BrCris, o mapeamento das fontes de dados a serem agregadas pelo Sistema, a implementação de provas de agregação dos recursos mapeados, a definição e realização de testes de serviços a serem disponibilizados. Entre as fontes agregadas, destacam-se em âmbito nacional, o OasisBr, a BDTD, a Plataforma Lattes, a Plataforma Sucupira e o Portal de Dados Abertos da CAPES. Já entre as fontes internacionais, destaque é dado ao OpenAIRE Research Graph, DOIBoost, Portal Wikidata e ao DOAJ.

A plataforma *LA Referencia* é flexível quanto ao modelo de dados adotado para armazenamento das informações. Para consulta externa, os dados coletados no âmbito do projeto BrCris serão também representados em um modelo semântico baseado na Ontologia VIVO, um modelo bastante utilizado internacionalmente por diversos CRISs. A representação em RDF com base em uma ontologia permite que o BrCris disponibilize Linked Open Data (LOD), tornando os dados não só abertos, mas também acessíveis e interoperáveis, seguindo, desta forma, as boas práticas disciplinadas pelos princípios FAIR.

Todas as ações de mapeio, coleta, transformação entre formatos e carga na plataforma *LA Referencia* são executadas por módulos desenvolvidos em Python, o qual foi denominado, Módulo BrCris, que se caracteriza como o orquestrador do Sistema. Dos dados organizados na plataforma *LA Referencia*, é possível indexar as informações agregadas no motor de busca Elasticsearch. Esta ferramenta é um mecanismo de busca de texto completo de código aberto escrito em Java que foi projetado para receber grandes volumes de dados além de ser distributivo e escalável.

Figura 5 – Entidades identificadas.

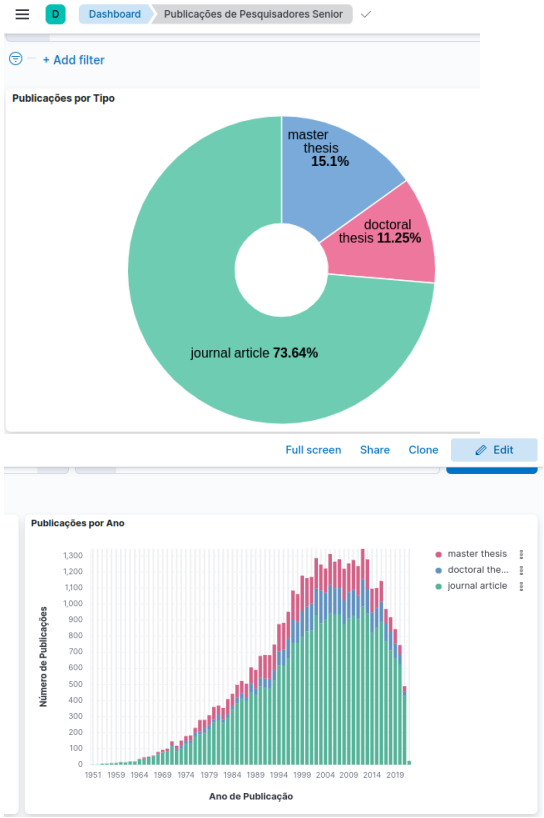


Dados da Pesquisa, 2022.

Para efeito de testes, já foram dadas cargas de dados e a partir destas cargas, foram disponibilizados os primeiros dashboards na ferramenta Kibana, plugada automaticamente aos índices do Elasticsearch, para que outros pesquisadores possam realizar análises dos dados importados sem a necessidade de softwares fechados ou que dependam de licenças de

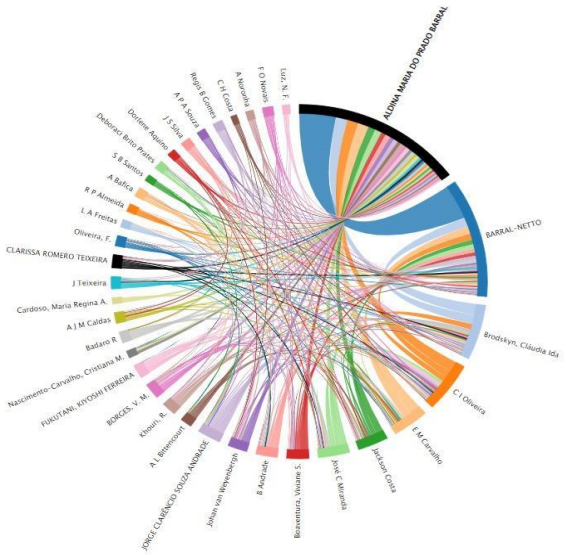
uso proprietárias ou pagas (ver exemplo nas Figuras 5, 6, 7 e 8).

Figura 6 – Dashboards de publicações.



Dados da Pesquisa, 2022.

Figura 7 – Redes de colaboração.



Dados da Pesquisa, 2022.

Figura 8 – Dashboards de patentes.

Dados da Pesquisa, 2022.

4 Considerações Finais

O BrCris se configura como um importante espaço de pesquisa e análise de dados. As informações agregadas e organizadas segundo um modelo de dados semântico, permitem a geração de serviços para diversos atores, nos contextos de gestão e pesquisa acadêmica, assim como na área de informação para a inovação, que pretende ser o alvo da proposta apresentada. O BrCris é uma iniciativa que coleta e enriquece dados de repositórios e bases de dados abertas pela *LA Referencia*, utilizando protocolos OAI-PMH e múltiplos formatos de dados em XML e JSON. A próxima etapa do projeto é a aplicação dos sistemas de recomendações pelas métricas que podem ser explanadas em cada conjunto de dados.

Referências

Kong, X. et al. Academic social networks: Modeling, analysis, mining and applications. Journal of Network and Computer Applications, v. 132, p. 86-103, 2019.

Pinto, A. L. et al. BrCris como um sistema de recomendação cintífico-tecnológica. In: Encontro Nacional de Pesquisa em Ciência da Informação, 2021.

Sivertsen, G. (2019) Developing Current Research Information Systems (CRIS) as data sources for studies of research. Springer handbook of science and technology indicators, p. 667-683.

PREVISÃO DE CRIMES EM VALÊNCIA: UMA ABORDAGEM MULTI-RÓTULO

CRIME PREDICTION IN VALENCIA: A MULTI-LABEL APPROACH

Luís Eduardo Freire da Câmara[1], Alexandre Rodrigues Loureiros[2], Jorge Mateu Mahiques[3], Flávio Miguel Varejão[4]

(1) UFES, Brasil, Espírito Santo, luis.camara@edu.ufes.br
(2) UFES, Brasil, Espirito Santo, arodrigues.ufes@gmail.com
(3) Universitat Jaume I, Espanha, Castellón, mateu@mat.uji.es
(4) UFES, Brasil, Espírito Santo, fmvarejao@gmail.com

Resumo

O aumento da criminalidade em áreas com alta concentração populacional torna as cidades uma das principais fontes de violência. Compreender as características de ocorrências criminosas e sua relação com a escala urbana transcende as questões acadêmicas e se torna central para a sociedade atual. Destarte, propomos uma abordagem de classificação multirrótulo para prever a ocorrência de crimes nas ruas da cidade de Valência, Espanha. Para tal proposta, utilizamos a base de dados criminais de Valência entre os anos de 2010 a 2020. Transformamos o problema de previsão de crimes nas ruas em um problema de classificação multirrótulo. Adicionalmente, aspectos meteorológicos e espaciais foram considerados para auxiliar na previsão dos crimes. Devido ao alto número de rótulos, utilizamos métodos de classificação extrema multirrótulo, em especial, o método Bonsai. Aplicando a métrica precisão@k, foi possível notar que o método Bonsai foi superior aos métodos multirrótulo tradicionais. A metodologia desenvolvida para o problema proporciona diversos avanços na área de previsão de crimes.

Palavras-chave: Aprendizado de máquina, previsão de crimes, multirrótulo.

Abstract

The increase in crime in areas with high population concentration makes cities one of the main sources of violence. Understanding the characteristics of criminal occurrences and their relationship with the urban scale transcends academic issues and becomes central to today's society. Thus, we propose a multi-label classification approach to predict crimes in the streets of Valencia, Spain. For this proposal, we used a crime database between the years 2010 to 2020. We transform a crime prediction into a multi-label problem. In addition, meteorological and spatial aspects were considered to assist in crime prediction. Due to the high number of labels, extreme multi-label methods were considered, especially the Bonsai method. Applying the precision@k metric to evaluate the models, it was possible to notice that the Bonsai method outperformed traditional approaches. The proposed methodology provides several advances in the area of crime prediction.

Keywords: Machine learning, crime prediction, multi-label.

1 Introdução

De acordo com a ONU (Organização das Nações Unidas), a população vem diminuindo sua taxa de crescimento populacional mundial desde 1950 (NAÇÕES UNIDAS BRASIL, 2022), no entanto, a elevação da população durante os anos gera diversos desafios de gestão de recursos, um deles sendo a segurança pública. Estudos mostram que existem diversos fatores que podem influenciar a taxa de insegurança da população. Em uma análise de agentes socioeconômicos e espaciais para taxa de assaltos, CHANG (2011) verificaram em seu estudo que houve uma taxa maior de crimes em locais onde apresentam mais casas e estabelecimentos comerciais, como também que o tipo de estrada, como um fator estrutural espacial, mostrou-se mais correlacionado com as taxas de invasão do que qualquer outro fator. Outros estudos sobre fatores espaciais foram tratados em ANSELIN et al. (2000), no que tange, a taxa de crimes em uma determinada localidade.

A ampla gama de fatores que influenciam as taxas de crimes, evidencia a importância do uso de métodos preditivos para auxiliar na gestão da segurança pública. De fato, recentes pesquisas focaram no uso de aprendizado de máquinas/estatística para criar ferramentas para auxiliar o combate ao crime. HOSSAIN (2020) investigaram o uso de uma metodologia utilizando técnicas de aprendizado supervisionado clássicos para

prever crimes. BOGOMOLOV (2014) propuseram um novo método de previsão de crimes geoespaciais a partir de múltiplas fontes de dados, em especial telefone móveis e dados demográficos. RODRIGUES (2010) utilizaram métodos de estatística espacial, como geoestatística e processos pontuais, para propor um sistema de vigilância para detectar aglomerados espaço-temporais emergentes na distribuição de homicídios na cidade brasileira de Belo Horizonte.

Este artigo propõe uma metodologia baseada em classificação multirrótulo utilizado dados temporais e espaciais dos crimes para prever tais delitos na cidade de Valência (Espanha), que passa por uma onda de crimes desde o início de 2021 (EPDATA, 2022). Portanto, há uma clara necessidade de pesquisa e desenvolvimento de tecnologias para encontrar soluções que ajudem a prevenir o aumento do número de casos de violência.

O restante deste artigo continua da seguinte forma: a Seção 2 descreve o objetivo proposto pelo artigo. A Seção 3 trata da seleção de técnicas, recursos e métodos de pesquisa. Nesta seção são apresentados a base de dados utilizada, os métodos multirrótulo, a descrição dos experimentos realizados e a métrica de avaliação. A Seção 4 apresenta os resultados obtidos. Por fim, a Seção 5 apresenta as considerações finais do artigo.

2 Objetivos

Este artigo investiga técnicas de classificação multirrótulo para propor uma metodologia para prever a ocorrência de crimes nas ruas da cidade de Valência, Espanha.

3 Procedimentos Metodológicos

Esta seção descreve a metodologia utilizada neste artigo. Todos os principais passos do método proposto estão apresentados no fluxograma da Figura 1. O método tem como entrada o conjunto de dados criminais, onde é atribuída uma abordagem multirrótulo aos dados. Com esta entrada, é feita uma divisão do conjunto para o treino do método preditivo. Antes de utilizar o modelo, o método é avaliado na base de teste. As seções a seguir descrevem os principais pontos do método proposto.

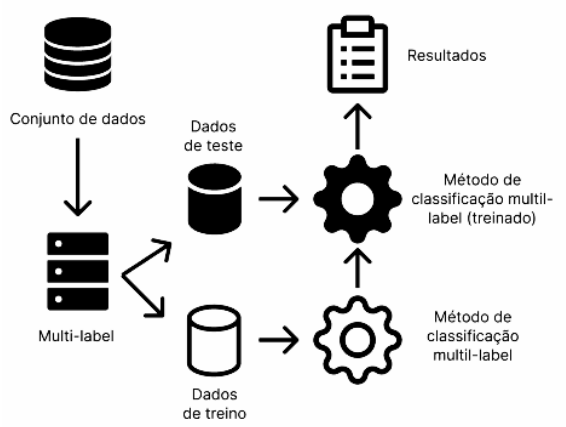


Figura 1 - Fluxograma com visão global da metodologia proposta.

3.1 Conjunto de dados

Originalmente, o conjunto de dados continha atividades criminosas em toda a cidade de Valência de 2010 a 2020. Neste conjunto de dados, haviam dados históricos dos crimes como latitude e longitude, a data, as distâncias de locais importantes envolvidos no crime, como bar, boate, caixa eletrônico, entre outras informações.

Para adicionar mais informações sobre o fato, foram adicionados dados meteorológicos extraídos da estação meteorológica da região. As características coletadas foram temperatura mínima, máxima e média, pressão mínima, máxima e média referentes a data do acontecido.

Todas as informações obtidas foram cedidas pelos serviços de dados abertos do governo.

3.2 Multirrótulo

O paradigma de classificação multirrótulo (*multi-label*) atribui a cada amostra um conjunto de rótulos alvo. Formalmente, para um denominando o espaço de características e $\omega = \{l_1, l_2, l_3, \ldots, l_n\}$ sendo um conjunto de rótulos. Então, uma amostra é definida como um par de vetores (x, y) onde x e y sendo uma combinação de rótulos, representado por um vetor binário em que $y = (y_1, y_2, \ldots, y_n)$, onde $y_i = 1$, apenas em amostras que sejam associadas a um rótulo l_i (MELLO; VAREJÃO; RODRIGUES, 2022).

No nosso contexto, foi utilizada a estratégia multirrótulo de forma que os dados

foram organizados como uma matriz de dias por rua, onde cada dia representa uma amostra e as ruas da cidade representam os rótulos. Cada amostra (dia) é associada a um ou mais rótulos (ruas) considerando a ocorrência de crimes naquele dia, isto é, cada amostra é associada às ruas que tiveram a ocorrência de 1 ou mais crimes naquele dia. Para conhecer a rua representativa do crime, foi utilizada a rua mais próxima da sua localização. Considerando o período de estudo (2010 a 2020), obtivemos ao todo 3.652 amostras (dias) e 2078 rótulos (ruas). Para compor o conjunto de características χ foram consideradas informações de 3 fontes: distância, meteorológicas e históricas. Maiores detalhes sobre as características são apresentados na Seção 3.3. Após a fase de treinamento, podemos usar o modelo para predizer as ruas com maiores probabilidades de ocorrência de crimes no dia seguinte (usar a informação de hoje para ver crimes de amanhã).

Devido ao alto número de rótulos (2078 ruas), optamos por utilizar métodos XMC (*Extreme Multi Label Classification*), onde são melhor explicados em SHEN et al. (2020), onde é desenvolvido um classificador que aponta para um subconjunto de rótulos mais relevantes de um conjunto muito grande de rótulos.

A principal qualidade da abordagem de algoritmos multirrótulo é sua capacidade de tratar dados esparsos. Como abordagem principal, utilizamos o algoritmo Bonsai, que generaliza o conceito de representação de rótulos em XMC através da representação do espaço de rótulos em árvores rasas (KHANDAGALE; XIAO; BABBAR, 2020). O Bonsai pode ser visto com uma extensão do método Parabel (PRABHU et al., 2018), que é muito usado em aplicações de XMC.

Para investigar esses algoritmos, foi utilizada a biblioteca Omikuji, que é uma implementação eficiente de "árvores de rótulos particionadas" e sua variante para XMC. Esta biblioteca oferece a possibilidade de treinar/testar algoritmos de XMC alterando hiperparâmetros. (DONG; SUOMINEN, 2022)

3.3 Organização de experimentos

Inicialmente, todas as características utilizadas neste problema multirrótulo foram divididas em 3 conjuntos. O primeiro é composto por características históricas, representando a quantidade de acidentes registrados anteriormente em cada rua. O segundo conjunto de características é composto por medidas meteorológicas diárias. Por fim, o terceiro conjunto usa informações espaciais, representando as distâncias médias, mínimas e máximas entre pontos de interesse da localização (boates, bares, etc) e os crimes.

As amostras foram distribuídas em 70% para treino e 30% para teste. A partição é sequencial para garantir a propriedade de continuidade dos dados históricos. Ou seja, os 70% primeiros dias são considerados amostra de treinamento e os 30% demais dias são utilizados para avaliar o modelo.

O método Bonsai foi treinado e avaliado para cada possível combinação de conjunto de características. Como forma de comparação, outros métodos multirrótulo foram testados considerando a combinação de conjuntos de características que apresentou melhor resultado no método Bonsai. Um dos métodos utilizados foi o *Binary Relevance* (BR), sendo uma das abordagens mais básicas para classificação multirrótulo, (ZHANG et al., 2018). A principal característica deste método é que ele ignora os relacionamentos entre rótulos. Além deste, o método XMC Parabel também foi treinado e avaliado para permitir comparação com o método Bonsai.

3.5 Métrica de avaliação

Para avaliar os resultados obtidos pelos algoritmos, a métrica precisão@k (P@k) foi utilizada. Esta métrica é amplamente utilizada em diversas aplicações de sistemas de recomendações e XMC, veja por exemplo JASINSKA-KOBUS et al. (2020), PRABHU et al. (2018), MITTAL et al. (2022). A Precisão@k é definida na Equação (1).

$$Precisão@k(\hat{y}, y) := \frac{1}{k} \sum_{l \in rank_k(\hat{y})} y_l \qquad (1)$$

A métrica Precisão@k é definida para um vetor de resultados previstos $\hat{y} \in \mathbb{R}^{\omega}$ e um vetor contendo os resultados reais

$y \in \{0, 1\}^{\omega}$. A função $rank_k(y)$ retorna os k maiores índices de $\hat{y}$ (KHANDAGALE; XIAO; BABBAR, 2020).

P@k é uma medida que corresponde ao número de resultados relevantes (rótulos positivos) que pertencem ao conjunto dos k rótulos com maiores probabilidades estimadas de ocorrência. Por exemplo, P@10 ou "precisão em 10" corresponde ao percentual de resultados relevantes (rótulos positivos) nos 10 rótulos com maiores probabilidades estimadas de ocorrência. Quanto maior for esta medida, mais assertivo é o modelo de previsão. Para este trabalho, consideramos k ∈ {*1, 3, 5*}.

4 Resultados

Esta seção apresenta os resultados obtidos nos experimentos descritos na seção 3.6.

A melhor combinação encontrada para o método Bonsai foi a com dados históricos e meteorológicos, sendo os resultados de P@k de 29.51, 23.59 e 21.37 para k sendo 1, 3 e 5 respectivamente. Como forma de comparação, os métodos Parabel e *Binary Relevance* também foram avaliados usando este mesmo conjunto de características. Todos os resultados do método Bonsai e os resultados dos métodos alternativos são exibidos na tabela do apêndice A. Como podemos observar na tabela, o subconjunto de variáveis espaciais apresentam um menor grau de explicação do problema.

Para melhor visualizar as diferenças entres os métodos nos diferentes valores de k, foi criado um gráfico de linha (Figura 2). Neste gráfico, é possível notar que os resultados do método Bonsai são superiores ao Parabel nas métricas P@1 e P@5. Ao passo que em P@3, o Parabel obteve desempenho ligeiramente superior ao Bonsai.

Adicionalmente, podemos perceber que o método BR apresentou resultados insatisfatórios e muito inferiores ao algoritmos de XMC. Uma explicação para isto é que o BR não considera correlação entre os rótulos, o que desfavorece seu resultado comparado com métodos XMC. Vale ressaltar que o esforço computacional necessário para obter os resultados do BR foram em média ≃ 80% maiores que para os métodos de XMC.

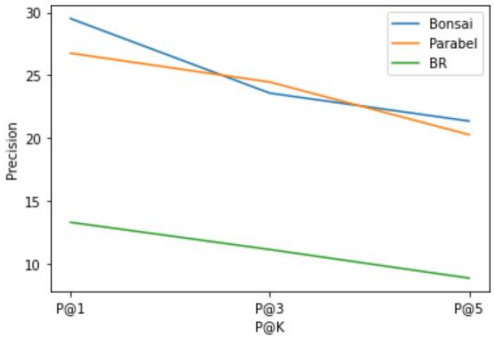


Figura 2 - Gráfico de comparação entre métodos (dados da pesquisa, 2022)

5 Considerações Finais

Este artigo transforma o problema de previsão de crimes em Valência, Espanha, em problema de classificação multirrótulo. Devido ao alto número de rótulos e a relação entre os mesmos, os métodos de XMC se mostraram superiores em comparação aos métodos tradicionais de multirrótulo.

Para facilitar a visualização dos resultados, criamos um gráfico onde as ruas são coloridas baseadas nas suas respectivas probabilidades de ocorrências de crimes. A Figura 3 apresenta uma região com ruas com alta probabilidade de ocorrência de crimes. O mapa completo da cidade pode ser visto no apêndice B.

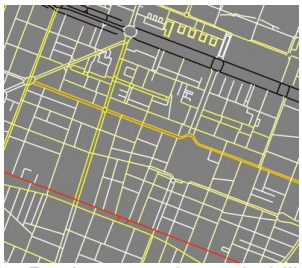


Figura 3 - Região com alta probabilidade de ocorrências criminosas

Metodologias utilizando a ideia de agrupamento entre ruas mais similares, o uso da análise de Componentes Principais (PCA) para redução dimensional, entre outras abordagens foram testadas. No entanto, todos resultados foram insatisfatórios e não reportados neste artigo.

Como trabalhos futuros destacamos o incremento de características no modelo para que as previsões tenham uma maior assertividade.

Referências

NAÇÕES UNIDAS BRASIL. População mundial chegará a 8 bilhões em novembro de 2022. [S.l] 2022. Disponível em <https://brasil.un.org/pt-br/189756-populacao-mundial-chegara-8-bilhoes-em-novembro-de-2022>. Acesso em: 05 set. 2022.

CHANG, Dongkuk. Social crime or spatial crime? Exploring the effects of social, economical, and spatial factors on burglary rates. Environment and behavior, v. 43, n. 1, p. 26-52, 2011

ANSELIN, Luc et al. Spatial analyses of crime. Criminal justice, v. 4, n. 2, p. 213-262, 2000.

EPDATA. Valencia - Crimen: asesinatos, robos, secuestros y otros delitos registrados en cada municipio. [S.l] [2022?]. Disponível em: <https://www.epdata.es/datos/crimen-asesinatos-robos-secuestros-otros-delitos-registrados-cada-municipio/6/valencia/7587> Acesso em: 8 set. 2022.

RODRIGUES, Alexandre Loureiros. Spatio-temporal models: low-rank approximation, inference and applications. 2010. Tese de Doutorado. Lancaster University.

HOSSAIN, Sohrab et al. Crime prediction using spatio-temporal data. In: International Conference on Computing Science, Communication and Security. Springer, Singapore, 2020. p. 277-289.

BOGOMOLOV, Andrey et al. Once upon a crime: towards crime prediction from demographics and mobile data. In: Proceedings of the 16th international conference on multimodal interaction. 2014. p. 427-434.

MELLO, Lucas Henrique Sousa; VAREJÃO, Flávio Miguel; RODRIGUES, Alexandre Loureiros. A Worst Case Analysis of Calibrated Label Ranking Multi-label Classification Method. Journal of Machine Learning Research, v. 23, n. 168, p. 1-30, 2022.

SHEN, Yanyao et al. Extreme multi-label classification from aggregated labels. In: International Conference on Machine Learning. PMLR, 2020. p. 8752-8762.

KHANDAGALE, Sujay; XIAO, Han; BABBAR, Rohit. Bonsai: diverse and shallow trees for extreme multi-label classification. Machine Learning, v. 109, n. 11, p. 2099-2119, 2020.

PRABHU, Yashoteja et al. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In: Proceedings of the 2018 World Wide Web Conference. 2018. p. 993-1002.

ZHANG, Min-Ling et al. Binary relevance for multi-label learning: an overview. Frontiers of Computer Science, v. 12, n. 2, p. 191-202, 2018.

JASINSKA-KOBUS, Kalina et al. Probabilistic label trees for extreme multi-label classification. arXiv preprint arXiv:2009.11218, 2020.

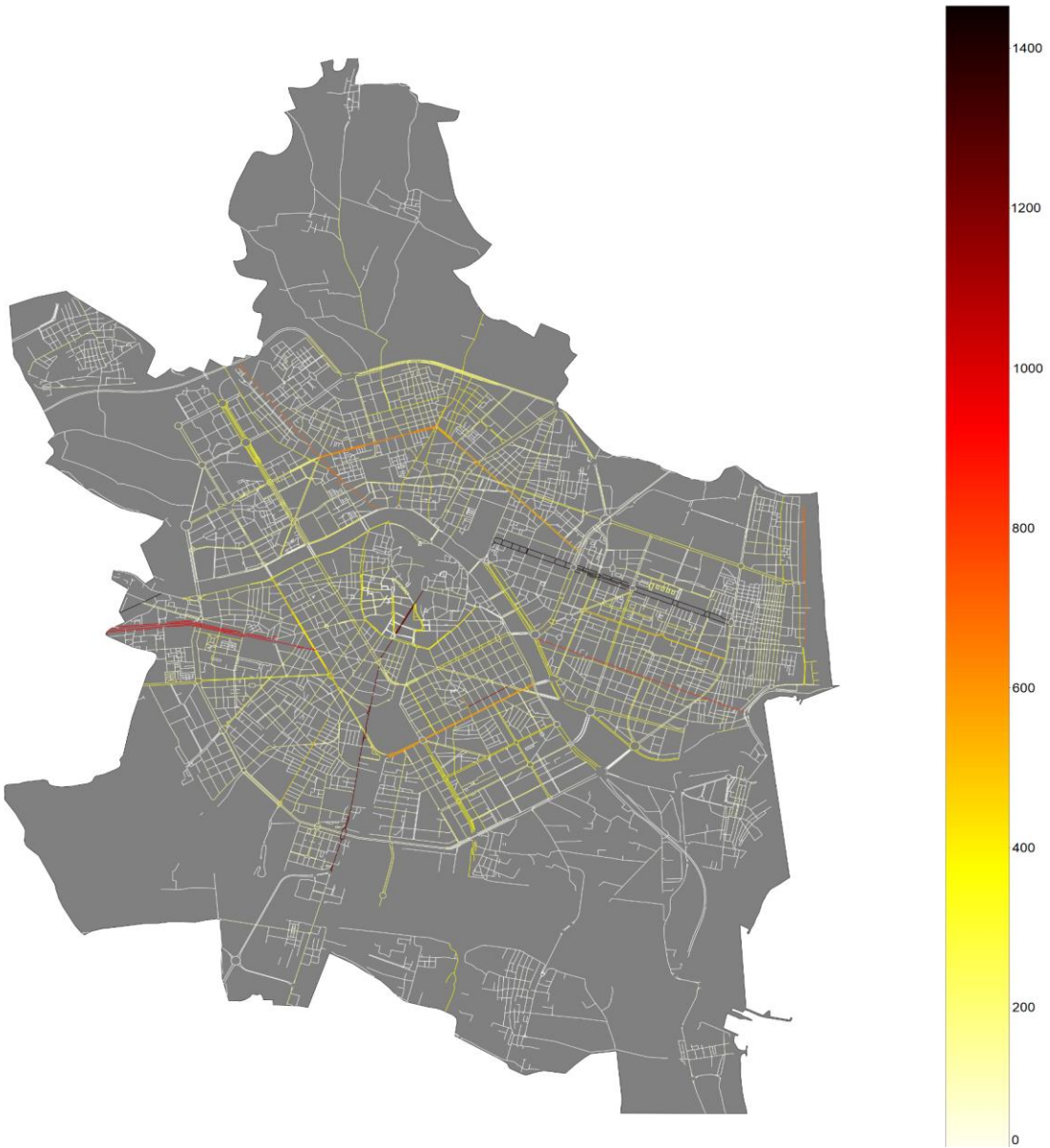
DONG, Tom; SUOMINEN, Osma. Omikuji. Disponível em https://github.com/tomtung/omikuji. Acesso em: 15 de set. 2022.

MITTAL, Anshul et al. Multi-modal Extreme Classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. p. 12393-12402.

Apêndice A - Tabela com resultados dos experimentos

Método	Dados Históricos	Dados Meteorológicos	Dados Espaciais	P@1	P@3	P@5
Bonsai	X	X	X	22.35	23.20	21.96
Bonsai	**X**	**X**		**29.51**	**23.59**	**21.37**
Bonsai	X		X	25.14	23.11	19.85
Bonsai		X	X	22.55	22.75	21.65
Bonsai	X			27.87	24.59	18.96
Bonsai		X		27.47	29.49	25.6
Bonsai			X	21.48	22.58	21.03
Parabel	X	X		26.76	24.47	20.29
BR	X	X		13.33	11.17	8.90

Apêndice B - Frequência de crimes por rua, (2010 - 2020)



PROCESSAMENTO DE LINGUAGEM NATURAL E MACHINE LEARNING COMO APARATO PARA A CATEGORIZAÇÃO DE ARTIGOS CIENTÍFICOS

NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING AS AN APPARATUS FOR THE CATEGORIZATION OF SCIENTIFIC PAPERS

Ananda Fernanda de Jesus[1], Maria Lígia Triques[2], José Eduardo Santarem Segundo[3], Ana Cristina de Albuquerque[4]

(1) Universidade Estadual Paulista (UNESP), Av. Hygino Muzzi Filho, 737 - Bairro Mirante, Marília/SP - CEP 17.525-900, af.jesus@unesp.br.

(2) Universidade Estadual de Londrina (UEL), Rodovia Celso Garcia Cid, PR-445, Km 380 - Campus Universitário, Londrina - PR, CEP 86057-970, mligia.triques@uel.br.

(3) Universidade Estadual Paulista (UNESP), Av. Hygino Muzzi Filho, 737 - Bairro Mirante, Marília/SP - CEP 17.525-900, Av. Hygino Muzzi Filho, 737 - Bairro Mirante, Marília/SP - CEP 17.525-900, santarem@usp.br.

(4) Universidade Estadual de Londrina (UEL), Rodovia Celso Garcia Cid, PR-445, Km 380 - Campus Universitário, Londrina - PR, CEP 86057-970, albuanati@uel.br.

Resumo

Este estudo objetiva verificar o potencial de aplicação de técnicas de Processamento de Linguagem Natural (PLN) e de *Machine Learning* (ML) na categorização temática de artigos científicos, por meio de categorias estabelecidas *a priori* e *a posteriori*. A partir da aplicação das técnicas de ML e PLN por meio de dois algoritmos de categorização (algoritmo de rede neural e algoritmo de clusterização hierárquica) em um *corpus* documental constituído de artigos científicos brasileiros sobre a temática "patrimônio cultural", desenvolve-se uma pesquisa aplicada, com resultados quantitativos e qualitativos derivados de dois procedimentos de teste: o primeiro utilizando algoritmo supervisionado, cuja categorização foi feita *a priori*; e o segundo, utilizando algoritmo não supervisionado, com a categorização feita *a posteriori.* Os resultados demonstram que para ambos os casos há a importância do detalhamento e rigor no pré-processamento dos dados e do tamanho e da representatividade da amostra escolhida. No caso supervisionado, conclui-se que quanto mais claras forem as características específicas de cada classe estabelecida *a priori* e quanto mais representativa for a mostra treino selecionada, maiores serão as chances de acerto do algoritmo. Já no caso não supervisionado, percebe-se que o algoritmo identifica de forma satisfatória o conteúdo dos documentos, permitindo inclusive identificar mais padrões de categorização que podem ser úteis às análises dos pesquisadores.

Palavras-chave: machine Learning; processamento de linguagem natural; algoritmo de rede neural; algoritmo de clusterização hierárquica; patrimônio cultural.

Abstract

This study aims to verify the potential of applying Natural Language Processing (NLP) and Machine Learning (ML) techniques in the thematic categorization of scientific articles, through categories established a priori and a posteriori. NLP and ML techniques are applied through two categorization algorithms (neural network algorithm and hierarchical clustering algorithm) in a documentary corpus composed of Brazilian scientific articles on the theme "cultural heritage", it is develop an applied research, with quantitative and qualitative results derived from two test procedures: the first using a supervised algorithm, which categorization was made a priori; and the second, using an unsupervised algorithm, with a posteriori categorization. The results demonstrate that, for both cases, there is an importance of detail and rigor in the pre-processing of the data and the chosen sample's size and representativeness. In the supervised case, it is concluded that the clearer the specific characteristics of each class established a priori and the more representative the selected training sample is, the greater the chances of success of the algorithm. In the unsupervised case, the algorithm satisfactorily identifies the documents content, even allowing the identification of more categorization patterns that can be useful for researchers' analyses.

Keywords: machine learning; natural language processing; neural network algorithm; hierarchical clustering algorithm; cultural heritage.

1 Introdução

A elaboração de categorias é um ato presente em diversas atividades cotidianas, desde a criação de categorias para organização dos espaços pessoais, das agendas de trabalho, até a estruturação das cidades e dos países.

A busca pela criação de categorias perpassa ainda o desenvolvimento e estabelecimento da Ciência da Informação, tendo em vista sua preocupação com a representação, "[...] origem, coleção, organização, armazenamento, recuperação, interpretação, transmissão, transformação, e utilização da informação." (BORKO, 1968, p.1). Destaca-se a necessidade de elaboração de categorias nas atividades de representação da informação e do conhecimento, visando sua posterior recuperação, tais como a catalogação, classificação e indexação.

Categorizar também é um ato importante para o desenvolvimento de pesquisas científicas, nas quais os resultados de estudos teóricos ou aplicados precisam ser agrupados e, assim, compreendidos enquanto um conjunto que possibilita a identificação de padrões e exceções, permitindo a geração de inferências. As categorias desses resultados podem ser estabelecidas *a priori*, ou seja, os resultados são agrupados em categorias criadas antes de sua análise, ou *a posteriori*, quando as categorias são elaboradas com base em padrões identificados nos resultados obtidos.

2 Objetivos

O presente estudo busca verificar o potencial de aplicação de técnicas de Processamento de Linguagem Natural (PLN) e de *Machine Learning* (ML) na automação do processo de categorização temática de artigos científicos, de forma a replicar as duas principais formas manuais de categorização: *priori* e a *posteriori* por meio de um recorte experimental.

O PLN pode ser definido como "[...] um conjunto de técnicas computacionais para a análise de textos em um ou mais níveis linguísticos, com o propósito de simular o processamento humano da língua". (FERNEDA, 2003, p. 82). Já ML é pautada na construção de agentes computacionais capazes de aprender com a experiência, com base na aplicação de técnicas estatísticas, em especial, por meio de algoritmos, visando a identificação de padrões e a realização de predições. A primeira etapa do processo de ML é o treinamento, que ocorre por meio da inclusão de um *corpus* (de dados ou de recursos informacionais), que permite que o algoritmo identifique quais variáveis levam a determinado resultado. (JORDAN; MITCHELL, 2015; CONEGLIAN, 2020).

A escolha do algoritmo ou conjunto de algoritmos a ser utilizado no processo de predição é contextual, dependendo das características do *corpus* e dos objetivos da atividade a serem realizadas, destacando-se dois grandes conjuntos de algoritmos:

O primeiro tipo, aprendizado supervisionado, utiliza dados para treinamento, cujo resultado é conhecido e explicitado para o algoritmo. Assim, o algoritmo conhece a solução e a partir dele e dos dados definirá quais são os aspectos que devem ser considerados para classificar algo em uma categoria.

No segundo tipo, aprendizado não-supervisionado, não há um resultado ou a solução desejada previamente, sendo o treino realizado então, com padrões estatísticos nos conjuntos de dados (CONEGLIAN, 2020, p.127).

Para evidenciar e discutir tais questões, propõe-se a aplicação das técnicas de ML e PLN em um corpus documental constituído de artigos científicos brasileiros que figurem em seus tópicos de estudo a expressão ou termo 'patrimônio cultural' como uma expressão explicitada ou definida em seu contexto de estudo.

Para refletir as principais formas manuais de categorização, o *corpus* selecionado foi submetido a algoritmos supervisionados e não supervisionados. Nesse cenário, os algoritmos supervisionados refletem a categorização *a priori*, em que se tem um conjunto de categorias conhecidas e busca-se "encaixar" novos documentos nessas categorias. Para a aplicação do algoritmo supervisionado o corpus selecionado foi previamente rotulado manualmente.

Já os algoritmos não-supervisionados foram aplicados visando refletir o processo de categorização *a posteriori*, quando os documentos observando são analisados e,

então, são criadas categorias. Para essa aplicação os documentos não foram previamente rotulados, observando-se o potencial do algoritmo na criação de novas categorias de análise.

Ressalta-se que para as duas situações, objetivou-se replicar, via aplicação de algoritmos, os procedimentos realizados de forma manual para categorização.

2 Procedimentos Metodológicos

A presente pesquisa é caracterizada como um estudo aplicado, com resultados quantitativos e qualitativos, com a finalidade de verificar o potencial de aplicação das técnicas de PLN e ML na categorização de resultados de um levantamento bibliográfico, tendo como recorte artigos científicos, publicados nacionalmente, em base temática da Ciência da Informação, a respeito da definição de 'patrimônio cultural'.

Para tanto, recorreu-se a um levantamento bibliográfico, qualitativo e exploratório da literatura científica, delimitado à produção nacional em português, utilizando a Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI), devido ao amplo espectro de documentos nacionais da Ciência da Informação que indexam.

A partir do uso do termo 'patrimônio cultural' na delimitação temporal de 2012 a maio 2022 (momento da coleta dos dados) como estratégia de busca, o *corpus* foi formado mediante à existência do termo em questão como descritor nas palavras-chaves das publicações e, posteriormente, identificado se havia a definição do termo em seu contexto de estudo. Ao final foram selecionados 46 artigos, cuja leitura permitiu identificar duas categorias: contexto de estudo relacionado ao meio digital (categoria A); contexto de estudo não relacionado ao meio digital (categoria B).

Esse *corpus* foi analisado tendo em vista tanto a possibilidade de sua categorização por meio de um algoritmo supervisionado, que executa suas funções utilizando as categorias estabelecidas *a priori* (A e B), como com a aplicação de um algoritmo não supervisionado, observando o potencial de sua aplicação na criação de novas categorias de análise.

Diante disso, a primeira etapa foi a de pré-processamento dos dados, que consiste em técnicas cujo objetivo é melhorar a qualidade dos dados para o posterior processamento, eliminando elementos que podem influenciar indevidamente o processo, criando resultados indesejados.

No caso deste estudo, empregou-se a divisão do conteúdo do texto em unidades menores (chamadas *tokens*), omitindo pontuação, processo conhecido como tokenização. Após isso, foram aplicadas opções de transformação dos dados de modo a garantir a padronização, como remoção de URLs e demais *links*, bem como uniformização em letra minúscula.

Em sequência, foram aplicados filtros, que permitem remover ou manter uma seleção de palavras, como a definição por idioma, no caso, português, dado o *corpus* de análise. Nesta fase, aplica-se principalmente o processo de identificação de *stopwords*, palavras tais como artigos e conectivos, que se repetem ao longo do texto, mas que não refletem o seu significado. Também foram excluídos os números, tais como páginas e anos, facilitando a visualização dos termos significativos.

A partir disso, gerou-se uma primeira nuvem de palavras, na qual foi possível identificar outros termos que também precisavam ser excluídos, tais como letras soltas e informações relativas ao periódico/evento do artigo, entre outros.

Para a avaliação da aplicabilidade em categorização utilizando categorias construídas, que foram rotulados manualmente *a priori*, realizou-se a etapa de teste e treino de um conjunto de algoritmos supervisionados, sendo considerados os algoritmos Rede Neural, KNN e *Random Forest*.

O processo do treino/teste foi realizado levando em consideração 80% do *corpus* total (37 dos 46 artigos), os outros 9 artigos foram reservados para uma validação, realizada aleatoriamente.

O procedimento de treino/teste teve o *corpus* (37 artigos) com uma aplicação de *K-fold cross validation*[1] *para 20 repetições, com*

[1] A validação cruzada *k-fold* destina-se a estimar a habilidade do modelo em novos dados.

divisão de 70% para treino e 30% para teste. Esses parâmetros foram os melhores encontrados após alguns testes.

Utilizou-se a métrica de acurácia para avaliar o resultado dos algoritmos. Apesar do pequeno *corpus*, o que normalmente não favorece um algoritmo de Rede Neural, ele teve melhor desempenho com uma acurácia de 88%, já o KNN atingiu 84% e o *Random Forest* com 82%.

Para a validação, foi aplicado o algoritmo Rede Neural, levando em consideração 20% dos artigos (9 dos 46 artigos). Tais documentos já haviam sido previamente rotulados pelos pesquisadores de forma manual para permitir a checagem dos erros e acertos, mas essa rotulação não foi indicada ao algoritmo.

Para a avaliação da aplicabilidade em categorias construídas *a posteriori* foi utilizado o processo não supervisionado, isto é, utilizando padrões estatísticos por meio de um algoritmo de clusterização. Para isso, foi retirada a *feature* que indicava categorização (*target*) dos textos selecionados.

Novamente foi repetido o pré-processamento para garantir a qualidade dos dados e, então, escolhido os parâmetros de frequência dos termos e dos documentos de forma que fosse calculada a importância de uma palavra em um documento em relação a uma coleção de documentos, e não somente as palavras de maior ocorrência total.

Posteriormente foram calculadas as métricas de distância no conjunto de dados utilizando-se as referências Euclidiana e Jaccard, resultando, assim, na aproximação entre os textos similares e, consequentemente, em sua categorização. Para a finalização do estudo, optou-se pela métrica Jaccard, pois obteve-se os melhores resultados de clusterização.

3 Resultados

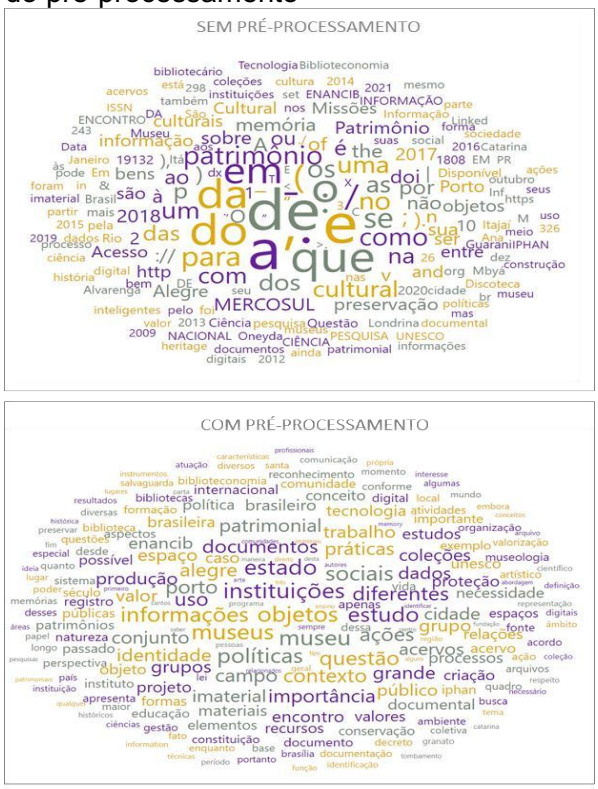
Como resultado da primeira etapa, com o pré-processamento já foi possível obter um panorama das discussões sobre patrimônio cultural, por meio da nuvem de palavras, que coloca em destaque os principais termos recorrentes no *corpus* teórico analisado.

Nesta etapa de geração da nuvem de palavras destaca-se a importância do processo de limpeza dos dados, o que fica evidente na Figura 1, em que se observa a nuvem de palavras antes e depois da remoção das *stopwords*.

Após a etapa de treino e teste já descrita, a etapa de validação teve apenas 1 dos 9 artigos classificados incorretamente, baseado na classificação manual, obtendo-se uma acurácia aproximada de 89%.

Compreendeu-se que a acurácia do algoritmo é influenciada por diversos fatores como o detalhamento e rigor no pré-processamento e na limpeza dos dados, o tamanho e a representatividade da amostra escolhida. Quanto mais claras forem as características específicas de cada classe estabelecida *a priori* e quanto mais representativa for a mostra treino selecionada, maiores serão as chances de acerto do algoritmo.

Figura 1 - Nuvem de palavras antes e depois do pré-processamento



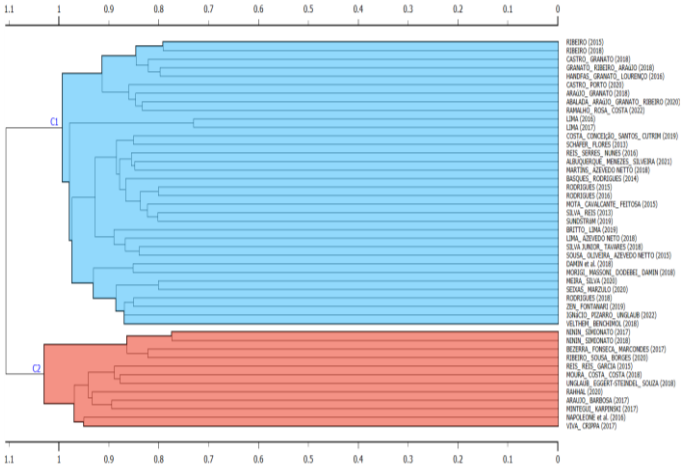
Fonte: Elaborado pelos autores.

Em relação ao processo não supervisionado, os resultados puderam ser verificados usando um algoritmo de clusterização hierárquica (*hierarchical clustering algorithm*) que permite a visualização dos documentos em função da aproximação ou distanciamento de seu conteúdo.

Isso se dá, pois, o algoritmo lê a métrica de distanciamento escolhida - no caso deste estudo, o índice Jaccard - e calcula uma matriz que categoriza os documentos que são mais similares, resultando em um dendrograma, tal como na Figura 3.

Com base nisso, o pesquisador pode combinar os itens de seu corpus a partir das métricas que lhe entregue melhores resultados.

Figura 3 - Dendograma do corpus de análise

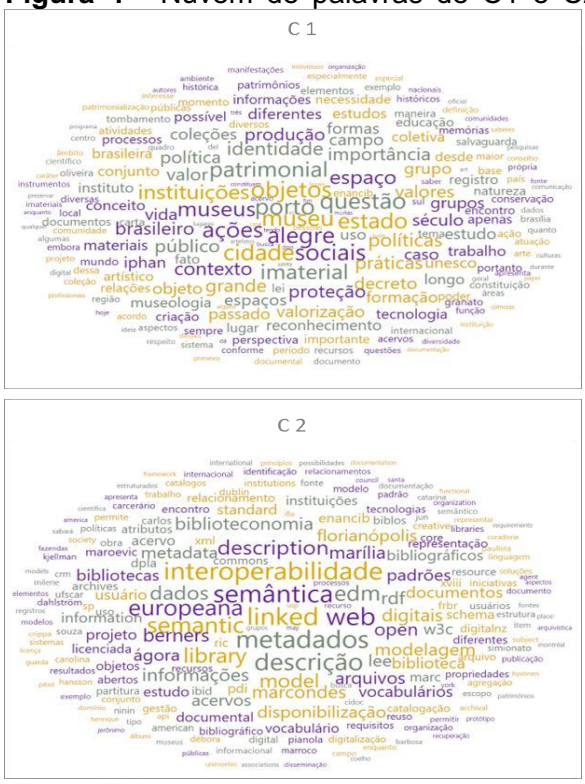


Fonte: Elaborado pelos autores.

Como o *corpus* analisado já havia passado por uma análise de conteúdo manual, resultando nas duas categorias (A e B) empregadas no primeiro procedimento (supervisionado), foi possível comparar se o segundo procedimento (não supervisionado) entregava resultados próximos ao que foi obtido manualmente.

Como é possível visualizar na Figura 3, o algoritmo gera um dendograma, ou seja, uma forma de visualização em formato de árvore com ramificações clusterizadas por similaridade. Ao marcar as duas categorias de maior nível na clusterização (C1 e C2), é possível gerar uma nuvem de palavras para cada uma (Figura 4), demonstrando que a primeira (C1) se aproxima da Categoria B identificada manualmente, isto é, referente a um contexto não vinculado ao digital. Já a segunda (C2), se aproxima da Categoria A identificada manualmente, o que corresponde ao contexto digital. Palavras como ‘museus’, ‘objetos’ e ‘sociais’, são as mais relevantes de C1, enquanto que ‘interoperabilidade’, ‘metadados’ e ‘web’ são os principais destaques da C2.

Figura 4 - Nuvem de palavras de C1 e C2



Fonte: Elaborado pelos autores.

Diante disso, verifica-se que o procedimento não supervisionado cria categorias que podem ajudar o pesquisador a decidir quais serão seus próprios parâmetros e suas próprias categorias de análise, isto é, categorias *a posteriori*, direcionando a leitura dos documentos e facilitando na compreensão do *corpus* em questão.

4 Considerações Finais

Entende-se que o estudo realizado colabora na compreensão do processo de análise de trabalhos científicos, sem ter aqui o objetivo de determinar ou obter bons resultados.

Considera-se também que não houve análises aprofundadas com experimentações de parâmetros até sua saturação nos algoritmos utilizados, o que se considera para um posterior aprofundamento da pesquisa, com a inclusão de um *corpus* mais robusto.

Em relação a predição de novos documentos com base em categorias obtidas *a priori*, conclui-se que ainda não é possível excluir a participação do pesquisador no

processo de categorização, entretanto o número de acertos faz com que a aplicação do processo seja relevante e possibilite imaginar um novo cenário, com o algoritmo atuando como uma pré-classificação, em que o pesquisador atuaria como validador, representando uma redução significativa de trabalho manual.

Outro papel importante do pesquisador no processo será o de seleção da amostra utilizada para treino do algoritmo e estabelecimento correto das características que diferenciam os conjuntos de dados e que permitem a criação das categorias, levando em consideração a influência desses quesitos na acurácia do algoritmo. Conclui-se ainda que o potencial de contribuição desse procedimento seria ampliado em análise de grandes volumes de documentos.

Já em relação a criação de categorias *a posteriori*, aplicando técnicas de PLN e ML, conclui-se que os resultados são mais promissores, tendo em vista que com base na aplicação desse procedimento é possível identificar novos padrões que poderiam passar despercebidos pelos próprios pesquisadores. Novamente o potencial do procedimento seria ampliado em um contexto de grandes volumes de documentos cuja análise aprofundada pelo pesquisador seria um processo longo e exaustivo, e em muitas situações inviável.

Tendo em vista a ampliação do potencial dessas técnicas na análise de grandes conjuntos de documentos, como estudos futuros, pretende-se ampliar a amostra utilizada e ainda realizar teste com outros algoritmos e esgotar (saturar) o uso de seus parâmetros com busca de melhores resultados. Pretende-se ainda verificar o potencial de aplicabilidade das técnicas em outras formas de análise de resultados de pesquisa científica, como na geração de métricas, em processos como Bibliometria.

Com base nas discussões apresentadas evidencia-se que as técnicas de Processamento de Linguagem Natural e de *Machine Learning* são promissoras para os processos de categorização de recursos informacionais, podendo contribuir assim com a redução do tempo despendido por profissionais especializados, incluindo os profissionais da informação no que tange às atividades de catalogação, classificação e indexação, bem como na categorização de resultados de pesquisas científicas.

Considera-se, portanto, que as técnicas podem contribuir tanto para a pré-categorização de novos recursos - quando as categorias desejadas já forem definidas -, como para a elaboração de novas categorias, permitindo assim identificação de padrões que poderiam passar despercebidos pelos pesquisadores ou evidenciando padrões já conhecidos.

Referências

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

CONEGLIAN, C. S. **Recuperação da Informação com abordagem semântica utilizando Linguagem Natural**: a Inteligência Artificial na Ciência da Informação. 2020. 194 f. Tese (Doutorado) - Curso de Programa de Pós-Graduação em Ciência da Informação, Universidade Estadual Paulista, Marília, 2020. Disponível em: https://repositorio.unesp.br/bitstream/handle/11449/193051/coneglian_cs_dr_mar.pdf?sequence=3&isAllowed=y. Acesso em: 08 set. 2022.

FERNEDA, E. **Recuperação de informação**: análise sobre a contribuição da ciência da computação para a ciência da informação. 2003. 137 f. Tese (Doutorado) - Curso de Programa de Pós-Graduação em Ciência da Informação, Universidade Estadual Paulista, Marília, 2003. Disponível em: https://teses.usp.br/teses/disponiveis/27/27143/tde-15032004-130230/fr.php. Acesso em: 08 set. 2022.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255-260, 2015. Disponível em: https://www.science.org/doi/abs/10.1126/science.aaa8415. Acesso em: 08 set. 2022.

PROCESSO SISTEMÁTICO FUNDAMENTADO EM MODELAGEM ONTOLÓGICA APLICADO À ESTRATIFICAÇÃO DE RISCO EM SAÚDE MENTAL PARA ANÁLISE QUALI-QUANTI

SYSTEMATIC PROCESS BASED ON ONTOLOGICAL MODELING APPLIED TO RISK STRATIFICATION IN MENTAL HEALTH FOR QUALI-QUANTITY ANALYSIS

Evaldo de Oliveira da Silva, Yuri Bento Marques, Marcello Peixoto Bax[1]
(1) Programa de Pós-Graduação em Ciência da Informação – UFMG, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, evaldosilva@ufmg.br, yuri.marques@ifnmg.edu.br, bax@eci.ufmg.br

Resumo
Transtornos mentais estão entre as causas de inaptidão no mundo. A ansiedade é um transtorno causado por trauma, estresse devido a uma doença ou a partir de outras desordens mentais. Estudos de saúde mental frequentemente utilizam estimativas e dados sociodemográficos e sobre psicopatologias. Utilizam-se instrumentos de avaliação que permitem tabular valores referentes aos sinais e sintomas e acompanhá-los na evolução do tratamento, como a estratificação de risco. Dados estruturados e não estruturados são gerados a partir desses instrumentos, o que dificulta as análises quantitativas e qualitativas, caso os sinais e sintomas sejam reavaliados ou se a eficácia e eficiência dos tratamentos serviram para mitigar o risco daquele paciente. Desta forma, este artigo apresenta um processo sistemático fundamentado na modelagem ontológica para análise quantitativa e qualitativa de dados da estratificação de risco em ansiedade, em conjunto com dados extraídos dos tratamentos psicoterápicos. O processo aborda o uso de ontologias e a técnica SDD (*Semantic Data Dictionary*) para compreensão e preparação dos dados. São gerados grafos de conhecimento que contém os dados anotados semanticamente. Esses grafos podem ser úteis para encontrar relações entre doenças, sintomas e tratamentos além de servir como ferramenta para formação de diagnóstico e auxílio na definição dos níveis de cuidados com o paciente.

Palavras-chave: Saúde Mental, Ontologia, Psicoterapia, Análise Quali-Quanti, Grafos de Conhecimento.

Abstract
Mental disorders are among the leading causes of disability in the world. Anxiety is a disorder caused by trauma, stress due to an illness, or from other mental disorders. Mental health studies often use estimates and socio-demographic and psychopathology data. Assessment tools are used to chart values for signs and symptoms and track them during treatment, such as risk stratification. Structured and non-structured data are generated from these instruments, which makes quantitative and qualitative analyses difficult, should signs and symptoms be reevaluated or if the effectiveness and efficiency of treatments have served to mitigate that patient's risk. Thus, this article presents a systematic process based on ontological modeling for the quantitative and qualitative analysis of risk stratification data in anxiety, together with data extracted from psychotherapeutic treatments. The process addresses the use of ontologies and the Semantic Data Dictionary (SDD) technique for data understanding and preparation. Knowledge graphs are generated and hold the semantically annotated data. These graphs can be useful for finding relationships between diseases, symptoms and treatments as well as serving as a tool for forming diagnoses and helping to define levels of patient care.

Keywords: Mental Health, Ontology, Psychotherapy, Quali-Quantity Analysis, Knowledge Graphs.

1. Introdução

Transtornos mentais acometem milhões de pessoas no mundo que sofrem de depressão, ansiedade e cometem suicídios (WHITEFORD, FERRARI e DEGENHARDT, 2016). A ansiedade é um transtorno mental causado por trauma, estresse devido a uma doença, ou ainda, a partir de outros transtornos mentais. Dados permitem análise da condição mental dos pacientes, por exemplo, instrumentos de avaliação contendo estruturas que permitem tabular escalas de valores referentes aos sinais e sintomas observados e os textos sobre a evolução dos tratamentos (MELLO, 2015; PAULA, 2019). Instrumentos de avaliação usados em estratificação de risco, revisam metas e orientam diferentes intervenções, conforme as necessidades dos pacientes (BRASIL, 2014).

A estratificação de riscos é um processo que utiliza instrumentos de avaliação no formato de questionário, que é aplicado para atribuir uma faixa de risco ao paciente com base em uma pontuação final. Equipes multiprofissionais utilizam o risco para estabelecer metas dos tratamentos dos planos de intervenção com o objetivo de mitigar o risco do paciente (MELLO, 2015; (PAULA, 2019; PARANÁ, 2021). Esses conjuntos de dados (estratificação e tratamentos), podem estar em formatos distintos (estruturados ou formatos textuais) dificultando as análises da eficácia dos tratamentos, se os riscos foram mitigados ou quais fatores que influenciam no tratamento. Além disso, estudos científicos podem ser desenvolvidos com base nos conjuntos de dados mencionados.

A *data science* refere-se à aplicação de métodos de obtenção de resultados científicos por meio da utilização de computação, e grande volume de dados (BERTHOLD *et al*., 2010; SCHRÖER, KRUSE e GÓMEZ, 2021). Possui um processo suportado pelas fases iniciais de compreensão e preparação de dados.

A compreensão de dados visa entender se existem dados suficientes para responder a diferentes questões de pesquisa (BERTHOLD *et al*., 2010). Alguns trabalhos utilizam ontologias como mecanismos para compreensão de dados atribuindo-lhes significados (semântica) (BRISSON e COLLARD, 2008; SVÁTEK, RAUCH e RALBOVSKÝ, 2005). Ademais, pesquisas apresentam o uso de ontologias para a organização do conhecimento na saúde mental (BRENAS, SHIN e SHABAN-NEJAD, 2019; CEUSTERS e SMITH, 2010; HASTINGS, 2012; YAMADA, 2020).

A preparação de dados abrange a seleção e mapeamento dos conjuntos de dados para suportar as técnicas de modelagem evitando análises enviesadas (PATON, 2019).

Dicionários de dados e metadados facilitam o compartilhamento, classificação e o mapeamento de dados na área da saúde e oferecem apoio para a preparação de dados (AHIMA, 2022). O ONC[1] tem desenvolvido projetos para promover padrões de metadados e dados de saúde, que suportam os princípios localizáveis, acessíveis, interoperáveis e reutilizáveis (ou FAIR[2]) (HEALTHIT, 2020; ZAYAS-CABÁN e WALD, 2020).

O uso de SDD (*Semantic Data Dictionary*) tem sido aplicado em consonância com os princípios FAIR para diferentes estudos científicos. Rashid et. al (2017) utilizam padrões de metadados para configurar a anotação semântica usando SDD. Os autores recomendam a utilização da ontologia de topo SIO (*Semanticscience Integrated Ontology*) que fornece propriedades para descrever os relacionamentos entre objetos e atributos como modelo de representação do conhecimento na área biomédica.

Diante dos diferentes tipos de dados gerados a partir dos instrumentos de avaliação e tratamento dos casos clínicos, a seguinte questão é apresentada: "Como compreender dados estruturados e não estruturados a fim de prepará-los para análise quali-quanti?". Desta forma, este artigo apresenta um processo sistemático fundamentado em modelagem ontológica aplicado à estratificação de risco em saúde mental e tratamentos psicoterápicos para análise quali-quanti.

O restante do artigo está organizado da forma que segue. A Seção 2 descreve a fundamentação teórica. A Seção 3 estabelece quais procedimentos metodológicos são utilizados para o processo proposto. A Seção 4 apresenta os resultados alcançados. A Seção 5 descreve as considerações finais.

2. Fundamentação teórica

2.1. Dicionário Semântico de Dados

Rashid et. al (2017) usam padrões de metadados para configurar a anotação semântica por um SDD. A anotação semântica proposta por Rashid et. al (2017) utiliza os seguintes documentos:

- *InfoSheet*: referências para descrição dos SDDs;

[1] Office of the National Coordinator for Health Information Technology

[2] Findability, Accessibility, Interoperability, e Reuse

- *Dictionary Mapping*: anotação semântica das colunas das coleções de dados;
- *CodeBook*: códigos correspondentes a conceitos de ontologia;
- *Code Mapping*: mapeamento de termos dos datasets que correspondem a conceitos existentes na ontologia;
- *TimeLine*: anotação de intervalos temporais;
- *Properties Table*: para fins de customizar a descrição por outras ontologias de topo.

A ferramenta sdd2rdf[3] interpreta o SDD e processa os dados, formando um grafo RDF. Para acessar os dados anotados, o sdd2rdf cria consultas no formato SPARQL. São geradas também regras SWRL[4] que auxiliam em novas inferências.

A técnica SDD é utilizada neste artigo visando explicitar os conceitos sobre os dados para gerar grafos de conhecimento dos dados anotados semanticamente. Os grafos serão utilizados para encontrar relações entre doenças, sintomas e tratamentos, servindo para auxiliar na definição dos níveis de cuidados com o paciente.

2.2. Análise qualitativa apoiada por software

Análises qualitativas podem ser apoiadas por softwares (CAQDAS – *Computer Assisted Qualitative Data Analysis*). A tarefa de *coding* permite definir ou categorizar os dados (textuais) que são analisados para diferentes objetivos de pesquisas. Essa tarefa utiliza de procedimentos que revelam temas embutidos nos dados (WILLIAMS e MOSER, 2019). CAQDASs organizam os *codings* em estruturas baseadas em desenhos qualitativos, que servem de base para codificar (anotar) os textos dos tratamentos psicoterápicos.

2.3. O framework FLAIR

Akbik *et al*. (2019) apresentam o FLAIR[5], um framework para processamento de linguagem natural (NLP, *Natural Language Processing*), utilizado para facilitar a classificação de texto. O FLAIR permite que o pesquisador aplique modelos de NLP para reconhecer NER (*Named Entity Recognition*), PoS (*Part-of-Speech Tagging*), com suporte especial para dados biomédicos, desambiguação e classificação de termos. NLP é importante no contexto deste trabalho, para extração de termos que evidenciem as ações terapêuticas, técnicas aplicadas para mitigação de riscos e termos biomédicos existentes nos tratamentos psicoterápicos.

2.4. Trabalhos Relacionados

Em Han e Stolffel (2011) os dados textuais são anotados com ontologias e as informações são recuperadas de diferentes maneiras a fim de permitir que pesquisadores possam aprender e inferir novos conhecimentos.

Kapiszewski e Karcher (2021) destacam que estudos científicos devem envolver a transparência da pesquisa qualitativa, anotação, exportação de códigos por meio de CAQDAS, bem como o compartilhamento de dados. Além disso, devem conter informações referentes à descrição da análise dos dados, qual método foi utilizado e se os *codings* foram feitos de forma indutiva ou dedutiva.

Hocker *et al.* (2021) discutem a troca de conhecimento dos dados analisados qualitativamente. De acordo com os autores, o *coding* deve fornecer uma documentação formal para suportar o compartilhamento dos dados. A ontologia QualiCO, desenvolvida pelos autores supracitados, visa preencher esta lacuna com o uso de metadados. Os metadados permitem que pesquisadores anotem informações sobre os *codings* a fim de enriquecer o conhecimento sobre os dados por eles categorizados (HOCKER *et al.*, 2021).

3. Processo sistemático fundamentado em modelagem ontológica para análise quali-quanti

O processo apresentado neste trabalho é fundamentado no uso de ontologias para compreender e preparar dados para análise quali-quanti. O processo é definido em 5 etapas a seguir:

[3] https://github.com/tetherless-world/SemanticDataDictionary

[4] https://www.w3.org/Submission/SWRL/

[5] https://github.com/zalandoresearch/flair

1. Modelagem ontológica. Criação/ajuste de ontologia de domínio (fundamentada em ontologias de topo) para formalização dos conceitos tratados no problema de pesquisa, incluindo a reutilização de ontologias consolidadas no domínio do problema.
2. *Coding*. O processo utiliza CAQDAS para análise qualitativa de dados não estruturados. O *coding* é construído a partir dos conceitos existentes na ontologia. Profissionais especialistas do domínio ou pesquisadores devem participar da modelagem ontológica com o objetivo de consensuar o conhecimento para análise qualitativa.
3. Execução de NLP. Ao terminar a análise qualitativa, os *codings* são extraídos para serem armazenados em *datasets* com os respectivos textos codificados. Então, NLP é executado em cada *coding* extraído, a fim de reconhecer elementos textuais, tais como, sujeito, ação, complemento da ação, doença e informações adicionais.
4. Unificação da estrutura de dados. Os dados reconhecidos por NLP são organizados em conjunto com dados quantitativos em um mesmo formato tabular, a fim de facilitar a anotação semântica.
5. Anotação Semântica. A anotação baseia-se em Rashid et. al (2017). Permite gerar o grafo RDF (*script sdd2rdf*) persistido no banco de dados *triplestore*. A ontologia formaliza o vocabulário. Após definir o *dataset* para anotar, os seguintes artefatos são gerados:

- *Dictionary Mapping* (DM). Cada linha do *DM* mapeia uma coluna do dataset, formalizando-a conceitualmente e também suas relações e proveniências.
- *CodeBook*. Permite a criação dos seguintes campos: Coluna (entidade a ser anotada), Código, Descrição e a Classe da Ontologia.
- Grafo RDF. Processamento dos artefatos "SDD + Dados" pelo *script sdd2rdf*, gerando o RDF e armazenando-o em um banco de dados *triplestore* para consulta posterior.

Os dados dos objetos mapeados pelo SDD são as colunas do próprio *dataset*. Porém, Rashid *et al*. (2017) afirmam que os objetos descritos no *dataset* podem encontrar-se ali explícita ou implicitamente. Ou seja, no mesmo *dataset* podem ser gerados novos atributos provenientes de outros objetos implicitamente representados . Estes objetos serão explicitados no SDD e formalizados no grafo final gerado (pelo script sdd2rdf), favorecendo a sua integração nos níveis conceituais (ou intencionais) mais abstratos do projeto.

4. Resultados

Descreve-se um processo sistemático fundamentado em modelagem ontológica para análise quali-quanti aplicada à estratificação de risco em saúde mental e tratamentos psicoterápicos (representado na Figura 1). Os dados da estratificação se encontram em formato tabular. As etapas do processo seguem da forma que são executadas:
3. Modelagem ontológica. A ontologia MHMO[6] *(Mental Health Management Ontology*) auxilia na gestão das redes de saúde mental e modela diferentes transtornos mentais.
4. *Coding*. Utiliza-se como CAQDAS a ferramenta Atlas-Ti[7] para análise qualitativa dos dados dos tratamentos psicoterápicos. O *coding* é construído a partir dos conceitos existentes na ontologia. Conceitos relacionados ao desenho qualitativo do caso clínico, tais como, aspectos cognitivos, aspectos comportamentais ou hipótese diagnóstica são usados para ajuste na MHMO. Os conceitos da ontologia utilizados para o *coding* são criados na ferramenta Atlas-Ti (Figura 1a). Os casos clínicos são analisados qualitativamente. O *coding* é extraído com os dados codificados.
5. O framework FLAIR é executado em cada *coding* extraído a fim de reconhecer elementos textuais, tais como, sujeito, ação, complemento da ação, doença e informações adicionais dos tratamentos, a fim evidenciar as ações terapêuticas.
6. Os dados da estratificação de riscos e os elementos textuais reconhecidos na etapa anterior são organizados em seus respectivos *datasets* (Figura 1b e Figura 1c).
7. Anotação Semântica. A Tabela 1 traz o *Codebook*, que descreve os dados categoriais do *dataset*: *QualitativeAnalysis*,

[6] https://bioportal.bioontology.org/ontologies/MHMO
[7] https://atlasti.com/

RiskLevel e *Gender*. O DM (Tabelas 2 e 3) mapeia para ontologias (Sio e MHMO) as seguintes características para tratamento de saúde mental: *Patient*, *Name*, *Subject*, *Action*, *ComplementOfActon*, *Disease* e *AditionalInformation*. Os grafos RDFs representam os fragmentos de conhecimento dos tratamentos psicoterápicos (Figura 1d).

5. Considerações Finais

O processo apresentado neste trabalho é fundamentado em modelagem ontológica e estabelece etapas para compreensão (com uso de ontologias) e preparação de dados para análise quali-quanti. Discute-se a aplicação do processo em dados da estratificação de risco em ansiedade e tratamentos psicoterápicos.

O trabalho utiliza ontologias para enriquecer o conhecimento, com o objetivo de auxiliar na compreensão dos dados. O processo é independente de uma ontologia específica, tanto para análise qualitativa, quanto para anotação semântica, indo além dos trabalhos correlatos citados (HAN e STOLFFEL, 2011; HOCKER *et al*., 2021). A preparação de dados, apoiada por SDD, reutiliza o conhecimento disponível na ontologia a fim mapear os dados por meio de templates de metadados.

Argumentou-se neste trabalho que a geração de grafos de conhecimento, com o uso da anotação semântica pelo SDD, pode abrir caminho para análise de dados quantitativos e qualitativos sobre eficácia dos tratamentos, ou quais fatores que influenciam no tratamento.

Como trabalhos futuros, espera-se aplicar o processo apresentado utilizando a ferramenta HaDatAc[8], visando a ingestão dos dados em uma infraestrutura que permita aquisições combinadas de dados e metadados semânticos. O uso do HaDatAc poderá ampliar a possibilidade da análise quali-quanti dos tratamentos em ansiedade a partir de variáveis harmonizadas em repositórios de dados semânticos.

Referências

AHIMA.Healthcare Data Governance . 2022. Disponível em: https://rb.gy/pcjbdh. Acesso em 03 de junho de 2022.

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019, June). FLAIR: An easy-to-use framework for state-of-the-art NLP. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations) (pp. 54-59).

BRISSON, Laurent; COLLARD, Martine. An ontology driven data mining process. In: International Conference on Enterprise Information Systems. 2008. p. 54-61.

BERTHOLD, M.R., Borgelt, C., H¨oppner, F., Klawonn, F.: Guide to intelligent data analysis: how to intelligently make sense of real data. Springer (2010).

BRASIL. Ministério da Saúde. Departamento de Atenção Básica. Cadernos de Atenção Básica: Núcleo de Apoio à Saúde da Família - Volume 1: ferramentas para a gestão e para o trabalho cotidiano. Brasília: Ministério da Saúde; 2014. Disponível em: http://bvsms.saude.gov.br/bvs/publicacoes/nucleo_apoio_saude_familia_cab39.pdf

BRENAS, Jon Hael; SHIN, Eun Kyong; SHABAN-NEJAD, Arash. Adverse childhood experiences ontology for mental health surveillance, research, and evaluation: advanced knowledge representation and semantic web techniques. JMIR mental health, v. 6, n. 5, p. e13498, 2019.

CEUSTERS, Werner; SMITH, Barry. Foundations for a realist ontology of mental disease. Journal of biomedical semantics, v. 1, n. 1, p. 1-23, 2010.

HASTINGS, Janna *et al*. Representing mental functioning: Ontologies for mental health and disease. 2012.

HEALTHIT. Advancing Health Data and Metadata Standards. 2020. Disponível em: https://www.healthit.gov/topic/scientific-initiatives/advancing-health-data-and-metadata-standards. Acesso em 03 de junho de 2022.

HAN, Dong, and STOLFFEL, Kilianl. "Ontology based qualitative case studies for

[8] https://www.hadatac.org/

sustainability research." In Proceedings of the AI for an Intelligent Planet, pp. 1-8. 2011.

HOCKER, J., BIPAT, T., MCDONALD, D. W., & Zachry, M. (2021). Developing an Ontology for Qualitative Coding Schemas - QualiCO. Disponível em: https://eera-ecer.de/ecer-programmes/conference/26/contribution/50889/. Acesso em 28 de jun de 2022.

KAPISZEWSKI, D., & KARCHER, S. (2021). Transparency in practice in qualitative research. PS: Political Science & Politics, 54(2), 285-291.

MELLO, Veronica de Pádua. Caminhos da educação em saúde na atenção básica: proposta de reorganização do grupo do parque. 2015. Tese de Doutorado. Universidade de São Paulo.

PAULA, George Luiz Costa de. Classificação de risco em saúde mental: implicações clínicas, éticas e sócio-políticas. 2019.

PARANÁ (Estado). Rede de Saúde Mental. Disponível em: https://www.saude.pr.gov.br/Pagina/Saude-Mental. Acesso em 09 de abr de 2021.

PATON, Norman. Automating data preparation: Can we? should we? must we?. In: 21st International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data. 2019.
PYLE, Dorian. Data preparation for data mining. Morgan Kaufmann, 1999.

RASHID, Sabbir M. et al. The Semantic Data Dictionary Approach to Data Annotation & Integration. In: SemSci@ ISWC. 2017. p. 47-54.

SCHRÖER, C., KRUSE, F., & GÓMEZ, J. M. (2021). A systematic literature review on applying CRISP-DM process model. Procedia Computer Science, 181, 526-534.

SVÁTEK, Vojtěch; RAUCH, Jan; RALBOVSKÝ, Martin. Ontology-enhanced association mining. In: Semantics, Web and mining. Springer, Berlin, Heidelberg, 2005. p. 163-179.

WHITEFORD, Harvey; FERRARI, Alize; DEGENHARDT, Louisa. Global burden of disease studies: implications for mental and substance use disorders. Health Affairs, v. 35, n. 6, p. 1114-1120, 2016.

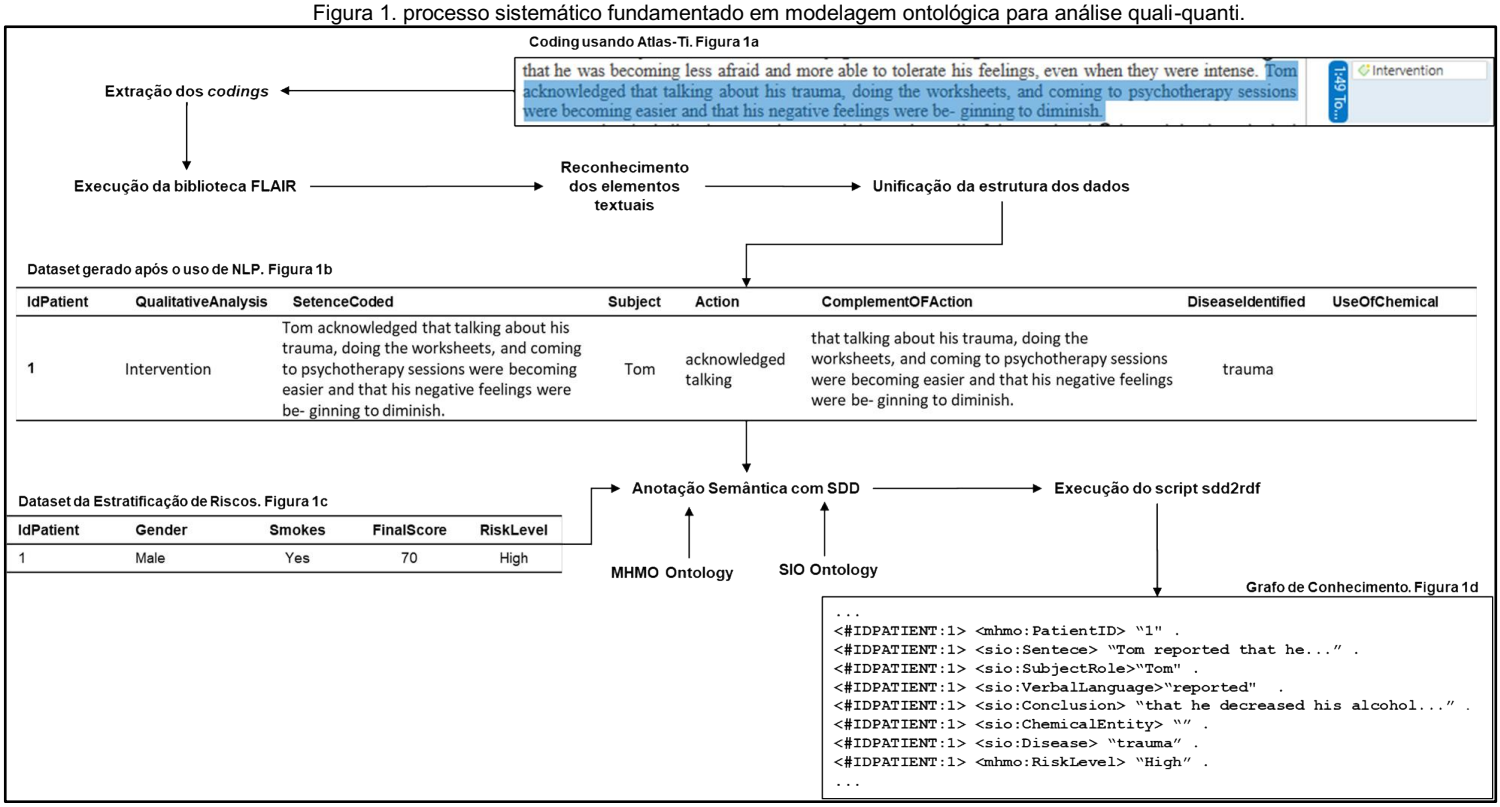
WILLIAMS, M., & MOSER, T. (2019). The art of coding and thematic exploration in qualitative research. International Management Review, 15(1), 45-55.

YAMADA, Diego Bettiol *et al*. Ontology-Based Inference for Supporting Clinical Decisions in Mental Health. In: International Conference on Computational Science. Springer, Cham, 2020. p. 363-375.

ZAYAS-CABÁN, Teresa, and Jonathan S. Wald. "Opportunities for the use of health information technology to support research." JAMIA open 3, no. 3 (2020): 321-325.

Apêndice A

Figura 1. processo sistemático fundamentado em modelagem ontológica para análise quali-quanti.



Fonte: Elaborada pelos autores.

Tabela 1. Codebook das .*QualitativeAnalysis*, *RiskLevel* e *Gender*.

Column	Code	Label	Class
Gender	Male	Gender of patient	mhmo:PatientGender
Gender	Female	Gender of patient	mhmo:PatientGender
Smokes	Yes	Patient smokes	mhmo:Smoking
Smokes	No	Patient smokes	mhmo:Smoking
RiskLevel	High	Risk level stratified	mhmo: RiskLevel
RiskLevel	Medium	Risk level stratified	mhmo: RiskLevel
RiskLevel	Low	Risk level stratified	mhmo: RiskLevel
UseOfChemicalSubstance	alcohol	Chemical Substance	mhmo:AlcoholDependence
Disease	anxiety	Mental Disorder	mhmo:AnxietyDisorder
QualitativeAnalysisOfClinicalCase	Diagnosis	Qualitative Analysis	mhmo:Diagnosis
QualitativeAnalysisOfClinicalCase	UseOfSubstance	Qualitative Analysis	mhmo:SubstanceDependence

Fonte: Elaborada pelos autores.

Tabela 2. Especificação do DM para dados explícitos.

Column	**Attibute**	**AttributeOf**	**Unit**
IDPatient	mhmo:PatientID	??patient	schema:Interger
SentenceCoded	sio:Sentence	??treatment	schema:Text
Subject	sio:SubjectRole	??treatment	schema:Text
Action	sio:VerbalLanguage	??treatment	schema:Text
ComplementOfAction	sio:Conclusion	??treatment	schema:Text
DiseaseIdentified	sio:Disease	??patient	schema:Text
UseOfChemical	sio:ChemicalEntity	??chemicalsubstance	schema:Text

Fonte: Elaborada pelos autores.

Tabela 3. Especificação do DM para dados implícitos.

Column	Entity	Relation	InRelationTo
??patient	sio:Human	mhmo:recevies	??treatment
??patient	sio:Human	mhmo:IsDiagnosedWith	mhmo:MentalDisorder
??patient	sio:Human	mhmo:takes	??chemicalsubstance
??patient	sio:Human	mhmo:hasRiskStratifiedIn	??risklevel
??patient	sio:Human	mhmo:uses	??chemicalsubstance

Fonte: Elaborada pelos autores.

PRODUÇÃO CIENTÍFICA EM CIÊNCIA DA INFORMAÇÃO: UTILIZANDO OS DADOS ABERTOS CAPES

SCIENTIFIC PRODUCTION IN INFORMATION SCIENCE: USING CAPES OPEN DATA

Patrícia Ofélia Pereira de Almeida
Universidade Estadual de Londrina, Londrina-PR, pereira@uel.br
Patrick Stacy Meyer
Universidade Estadual de Londrina, Londrina-PR, patrick.enzo.meyer@gmail.com

Resumo
A produção científica é um requisito fundamental para pesquisadores, e no âmbito da pós-graduação stricto sensu faz parte dos requisitos necessários para a obtenção do título de mestre ou doutor. O presente estudo tem como objetivo descrever alguns aspectos dos metadados do Catálogo de Teses e Dissertações disponibilizados pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, no que se refere à sua estrutura, e aos processos de coleta, limpeza e utilização para fins de pesquisa e análise, e ainda apresentar um panorama da produção em Ciência da Informação. Tem características quantitativa, descritiva e analítica. Foram coletados os conjuntos de dados do Catálogo de Teses e Dissertações – Brasil, referentes aos anos de 2017 a 2020. Como resultados, observou-se que as produções são mais recorrentes na região Sudeste, sendo que a maioria absoluta se refere a dissertações, e em seguida às teses. Houve uma queda no número de produções no ano de 2020.

Palavras-chave: Dados abertos; Ciência de dados; CAPES; Programas de pós-graduação em Ciência da Informação; Dissertações e teses.

Abstract
Scientific production is a fundamental requirement for researchers, and within the *stricto sensu* graduate program it is part of the requirements for obtaining a master's or doctorate degree. The present study aims to describe some aspects of the metadata of the Theses and Dissertations Catalog made available by the Coordination for the Improvement of Higher Education Personnel, in terms of its structure, and the processes of collection, cleaning and use for research and analysis, and also to present an overview of the production in Information Science. It has quantitative, descriptive and analytical characteristics. Datasets from the Catalog of Theses and Dissertations - Brazil were collected, referring to the years 2017 to 2020. As a result, it was observed that the productions are more recurrent in the Southeast region, with the absolute majority referring to dissertations, and then the theses. There was a drop in the number of productions in the year 2020.

Keywords: Open data; Data Science; CAPES; Postgraduate programs in Information Science; Dissertations and theses.

1 Introdução

A produção científica é um requisito fundamental para pesquisadores, pois é uma forma de divulgarem os resultados parciais e finais de suas investigações científicas, e obter a avaliação e o reconhecimento de seus pares.

No âmbito da pós-graduação *stricto sensu*, a publicação de artigos em periódicos, eventos, e outras formas de produção científica faz parte dos requisitos necessários para a obtenção do título de mestre ou doutor.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) coleta e disponibiliza os dados referentes à produção científica proveniente da pós-graduação no Brasil, de forma que é possível obter, tratar e analisar tais dados sob diversas perspectivas.

O presente estudo tem como objetivo descrever alguns aspectos dos metadados do Catálogo de Teses e Dissertações disponibilizados pela CAPES, no que se refere à sua estrutura, e aos processos de coleta, limpeza e utilização para fins de pesquisa e análise, e ainda apresentar um panorama da produção em Ciência da Informação nos respectivos registros.

2 Referencial Teórico

Trabalhar com um elevado volume de dados tem sido uma preocupação constante, principalmente a partir da década de 60, quando as indústrias, as pesquisas, e a comunicação científica começaram a evoluir

de forma mais ativa e com maior dimensão quantitativa de resultados.

A Ciência da Informação tem se preocupado com o tratamento de dados, no intuito de transformá-los em insumo de valor para a tomada de decisão, em especial depois que Borko (1968) definiu o escopo de estudo da Área.

Nesse sentido, Donoho (2015) afirma que a mais de 50 anos foi detectada a necessidade de trabalhar e aprender com dados, a fim de estabelecer novos métodos de uso da estatística. Porém, continua o autor, faz somente cerca de uma década que as principais universidades têm investido nos programas de ciência de dados.

Para Grus (2021, p. 18), o cientista de dados é um profissional que detém conhecimentos acerca de estatística e computação, e utiliza suas habilidades para extrair conhecimento de dados desorganizados.

Nesse sentido, pode-se dizer que o *Data Science*, ou Ciência de Dados, consiste em utilizar recursos da computação para tratar/organizar uma grande quantidade de dados, aplicando modelos matemáticos e estatísticos, de forma que os resultados sejam sintetizados e a análise dos dados se torne possível. Nessa direção, Coneglian, Santarém Segundo e Sant'ana, (2017) consideram que a análise de dados permite detectar padrões que, ao serem modelados, se tornam informações que dão suporte ao processo de tomada de decisão.

A CAPES (2022) apresenta que o objetivo da avaliação da pós-graduação *stricto sensu* no Brasil consiste na certificação da qualidade, assim como identificar discrepâncias regionais e de áreas estratégicas do conhecimento. Nesse sentido, identificar padrões de produção científica (ou a falta deles) no âmbito de um programa, Instituições de Ensino Superior (IES), regiões ou mesmo de forma global, é uma forma de diagnosticar problemas que podem estar prejudicando o rendimento dos programas, ou mesmo boas práticas que estejam elevando a produtividade.

Dessa forma, os dados disponibilizados pela CAPES são uma valiosa fonte de informação, que precisam ser tratados para que possam fornecer padrões e perspectivas acerca dos programas de pós-graduação *stricto sensu*, das áreas do conhecimento, e diversos outros aspectos possíveis.

Os metadados "crus" fornecidos pela CAPES são um apanhado de letras, números e símbolos sem sentido semântico, mas, se tratados, constituem um rico estoque de informações que podem ser modeladas para finalidades específicas.

3 Procedimentos Metodológicos

O presente estudo tem características quantitativa, descritiva e analítica, visto que pretende apresentar e agrupar um volumoso conjunto de metadados.

Para atingir o objetivo proposto, foram coletados os conjuntos de dados do Catálogo de Teses e Dissertações – Brasil[1], referentes ao quadriênio de 2017 a 2020.

Foram capturados os arquivos em formato CSV, de forma que pudesse ser mais facilmente utilizado pelo software Microsoft Excel, considerando ser um recurso que habitualmente está disponível em uma grande quantidade de computadores pessoais, além de ser fácil encontrar na Internet os tutoriais para usos de diversas aplicações. Obviamente existem outros *softwares* específicos para o tratamento de dados, que podem oferecer resultados mais rápidos e mais precisos, contudo a menor disponibilidade de tutoriais e a necessidade de treinamento do zero foram considerados como obstáculos, o que demanda de maior tempo de dedicação para ser superado.

Para a apresentação do panorama da produção em Ciência da Informação nos respectivos registros, foram selecionados os metadados categorizados com a respectiva área do conhecimento, com a finalidade de quantificação e apresentação dos resultados.

4 Resultados

Os arquivos recuperados (.csv) apresentaram em média o tamanho de 445 MB e 87.599 registros. Apesar de serem um pouco pesados, a estrutura dos dados em planilhas se mostrou de fácil entendimento e modelagem.

A Tabela 1 apresenta o ano, a quantidade de IES, o número de programas de Pós-Graduação e o total de produções registradas em cada conjunto de dados coletados.

Tabela 1 – Dados da produção da pós-graduação no Brasil – 2017 a 2020.

[1] https://dadosabertos.capes.gov.br/dataset/2017-2020-catalogo-de-teses-e-dissertacoes-da-capes.

ANO	IES	Programas	Produção	Cresc. anual (%)
2017	424	1.712	85.310	-
2018	441	1.788	90.469	6%
2019	453	1.828	94.503	4%
2020	466	1.819	80.114	-15%
Total			**350.396**	

Fonte: Dados da pesquisa.

Pode-se observar que o número geral de produções teve um crescimento gradativo de 2017 a 2019, mas regrediu bruscamente em 2020. É possível que essa queda se justifique pelo distanciamento social causado pela pandemia da Covid-19, cujos reflexos afetaram todos os setores e, portanto, com a educação não foi diferente. O corte de bolsas e a impossibilidade de aulas presenciais, dentre outros aspectos, foram prejudiciais para a manutenção do calendário acadêmico. Apenas com os dados do quadriênio 2021/2024 será possível analisar os reflexos quantitativos que o período causou para a pós-graduação no Brasil.

Os dados disponibilizados pela CAPES estão estruturados em 58 campos, os quais podem ser códigos numéricos ou textuais. Não foi possível localizar uma tabela com as siglas utilizadas para identificação das colunas de dados, e em alguns casos foi necessário recorrer ao conteúdo do campo para identificar seu teor. O sistema categoriza os campos por prefixos (AN, CD, DH, DS, DT, ID, IN, NM, NR, SG).

Foi possível identificar que os campos apresentam dados dos documentos no que se refere à representação descritiva (autor, título, número de páginas etc.), temática (palavras-chave, resumo etc.), vínculo (Instituição, programa, projeto etc.), área (área do conhecimento, linha de pesquisa etc.), pessoal (categoria, titulação etc.) e de data.

Observou-se que os dados permitem uma variada gama de recortes para análises quantitativas ou qualitativas. Contudo, para o objetivo de demonstrar o panorama da produção em Ciência da Informação, optou-se por descartar alguns campos que não se mostraram representativos para esta finalidade, tornando o arquivo dos metadados mais leve e maleável. Nesse sentido, foram mantidos os dados básicos de identificação dos documentos, o que resultou em 17 campos.

Do total geral de 350.396 registros de documentos recuperados, foram selecionados os da Área do Conhecimento Ciência da Informação, e organizados em um novo arquivo. Os dados foram agrupados aplicando os recursos de tabela dinâmica, com o intuito de apresentar a representação visual da informação. Nas 21 IES identificadas, constam cadastradas um total de 1520 produções em CI (Tabela 2).

Tabela 2 – Produção da pós-graduação em Ciência da Informação no Brasil por instituição/ano – 2017 a 2020.

REGIÃO/ IES	2017	2018	2019	2020	Total
Centro-Oeste	**27**	**29**	**26**	**21**	**103**
UNB	27	29	26	21	103
Nordeste	**61**	**89**	**129**	**93**	**372**
FUFSE	-	-	15	18	33
UFBA	9	23	21	5	58
UFC		15	9	7	31
UFCA	3	13	16	18	50
UFPB-JP	25	12	39	25	101
UFPE	16	17	13	11	57
UFRN	8	9	16	9	42
Norte	-	**1**	**9**	**16**	**26**
UFPA	-	1	9	16	26
Sudeste	**163**	**232**	**228**	**185**	**808**
FCRB	-	11	13	11	35
FUMEC	-	28	19	24	71
UFF	9	10	23	15	57
UFMG	37	36	46	44	163
UFRJ	19	32	35	28	114
UFSCAR	-	9	10	10	29
UNESP-MAR	33	43	48	34	158
UNIRIO	49	27	20	9	105
USP	16	36	14	10	76
Sul	**59**	**47**	**53**	**52**	**211**
UDESC	13	15	13	12	53

UEL	19	11	10	13	53
UFSC	27	21	30	27	105
Total CI	**310**	**398**	**445**	**367**	**1.520**
% do total geral	**0,36**	**0,44**	**0,47**	**0,46**	**0,43**

Fonte: Dados da pesquisa.

Apesar da redução de produções registradas em 2020, a Tabela 2 demonstra que CI manteve um percentual anual de produções próximo da média dos anos anteriores e acima da média total, de forma que não se mostrou prejudicada em maior medida do que as outras áreas.

Também é possível observar que a UFMG e a UNESP-MAR são as instituições com maior número de produções no quadriênio (10,8% e 10,4%, respectivamente). Em seguida aparecem a UFRJ (7,5%), a UNIRIO e a UFSC (6,91%), a UNB (6,78%) e a UFPB-JP (6,64%). As demais instituições somaram 44,1%.

A maior produção em CI está na região Sudeste, que detém mais da metade do total (53,2%), onde também se concentra um maior número de instituições com programas de pós-graduação na Área. O Nordeste aparece em segundo lugar (24,5%), seguido do Sul (13,9%), Centro-Oeste (6,8%) e Norte (1,7%).

Pode-se observar também que a redução na produção foi linear em 2020, atingiu a maioria das instituições, apenas seis IES mantiveram ou aumentaram o número de defesas (FUFSE, UFCA, UFPA, FUMEC, UFSCAR e UEL).

A Tabela 3 apresenta a produção da pós-graduação em Ciência da Informação no Brasil por estado.

Tabela 3 – Produção da pós-graduação em Ciência da Informação no Brasil por estado – 2017 a 2020.

REGIÃO/UF	2017	2018	2019	2020	Total
Centro-Oeste	**27**	**29**	**26**	**21**	**103**
DF	27	29	26	21	103
Nordeste	**61**	**89**	**129**	**93**	**372**
BA	9	23	21	5	58
CE	3	28	25	25	81
PB	25	12	39	25	101
PE	16	17	13	11	57
RN	8	9	16	9	42
SE	-	-	15	18	33
Norte	-	**1**	**9**	**16**	**26**
PA	-	1	9	16	26
Sudeste	**163**	**232**	**228**	**185**	**808**
MG	37	64	65	68	234
RJ	77	80	91	63	311
SP	49	88	72	54	263
Sul	**59**	**47**	**53**	**52**	**211**
PR	19	11	10	13	53
SC	40	36	43	39	158
Total	**310**	**398**	**445**	**367**	**1.520**

Fonte: Dados da pesquisa.

Observa-se que o Sudeste é a região mais produtiva, na qual as instituições estão de certa forma equilibradas no número de produções registradas. Contudo, destaca-se o estado do Rio de Janeiro com um acumulado de produções mais significativo.

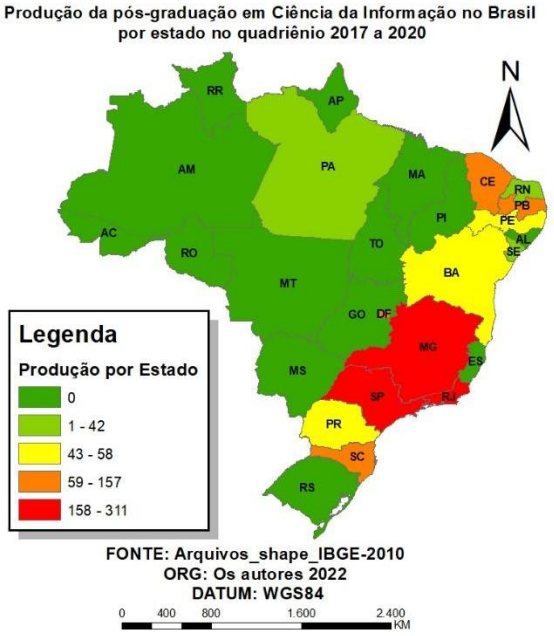
Esses dados podem ser visualizados na Figura 1, que destaca gradativamente a produção da pós-graduação *Stricto sensu* em CI por estado.

A representação visual proporcionada pelo mapa permite visualizar com maior destaque onde estão as IES mais produtivas em CI.

Na Tabela 4 estão apresentados os dados da CI por tipo de produção.

Embora tenham sido identificadas mais de 20 tipos de produções no contexto geral dos dados coletados, na área de Ciência da Informação constatou-se a presença de somente seis tipos. As dissertações são significativamente a maioria absoluta das produções recuperadas. Em seguida figuram as teses que, embora estejam em menor número, ainda são mais representativas que os demais tipos de produção.

Figura 1 – Produção da pós-graduação em Ciência da Informação no Brasil por estado – 2017 a 2020.



Fonte: Dados da pesquisa.

Tabela 4 – Produção da pós-graduação em Ciência da Informação no Brasil por tipo – 2017 a 2020.

TIPO	2017	2018	2019	2020	Total
Dissertação	227	296	310	259	1.092
Tese	73	93	121	96	383
Produto, proced. ou técnica	10	9	14	4	37
Projeto técnico	-	-	-	6	6
Editoria	-	-	-	1	1
Relatório final de pesquisa	-	-	-	1	1
Total	**310**	**398**	**445**	**367**	**1.520**

Fonte: Dados da pesquisa.

Ao todo, somaram-se 383 teses de doutorado, 802 dissertações de mestrado acadêmico, e no mestrado profissional são 290 dissertações e 45 produções de outros tipos. Identificou-se ainda 406 orientadores, contudo, apenas 62 (15,3%) estavam presentes nos quatro anos de dados coletados, e apenas sete (1,7%) orientaram 10 ou mais trabalhos defendidos. Cada orientador acompanhou de 1 a 15 pós-graduandos, perfazendo uma média de 3,8 defesas por orientador no período analisado.

5 Considerações Finais

Com base no exposto, foi possível utilizar os dados abertos da Capes para a realização da pesquisa, na qual identificou-se que houve uma queda da produção em 2020. Considerando que as dissertações lideram a maioria absoluta das produções, que a capacitação em nível de mestrado habitualmente ocorre no prazo de dois anos, associada à situação de pandemia que teve início no final do ano de 2019, pode-se supor que houve um atraso nas defesas, e que possivelmente o quantitativo que diminuiu em 2020 irá figurar no relatório CAPES de 2021.

Constatou-se também que a região Sudeste é um polo no que se refere à produção em CI, sendo que ali figuram também grandes centros de pesquisa e inovação tecnológica, mas que também sofreu uma queda em 2020.

Por fim, reitera-se a ideia de que a análise de dados permite uma série de perspectivas acerca daquilo que representam, e que os dados disponibilizados pela CAPES podem proporcionar inúmeras outras análises de cunho qualitativo.

Referências

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

CONEGLIAN, C. S.; SANTAREM SEGUNDO, J. E.; SANT'ANA, R. C. G. Big Data: fatores potencialmente discriminatórios em análise de dados. **Em Questão**, Porto Alegre, v. 23, n. 1, p. 62–86, 2017. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/62122. Acesso em: 2 set. 2022.

COORDENAÇÃO de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). **Dados abertos**. Brasília (DF): CAPES, 2022. Disponível em: https://dadosabertos.capes.gov.br. Acesso em: 05 set. 2022.

DONOHO, David. 50 Years of Data Science, **Journal of Computational and Graphical Statistics**, v. 26, n. 4, p. 745-766, 2017. Disponível em: https://www.tandfonline.com/doi/full/10.1080/10618600.2017.1384734. Acesso em: 17 set. 2022.

GRUS, Joel. **Data Science do zero**. 2. ed. Rio de Janeiro: Editora Alta Books, 2021.

QUALIFICAÇÃO DE REPOSITÓRIOS DE DADOS E DE PUBLICAÇÕES: UMA PROPOSTA DE CRITÉRIOS ALINHADA À CIÊNCIA ABERTA

QUALIFICATION OF DATA REPOSITERS AND PUBLICATIONS: A PROPOSAL OF GUIDELINES IN SUPPORT OF OPEN SCIENCE

Priscila Machado Borges Sena[1], Tatyane Guedes Martins da Silva[2], Juliana Araujo Gomes de Sousa[3], Washington Luís Ribeiro de Carvalho Segundo[4], Bianca Amaro[5]

(1) Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), SAUS Quadra 5 - Lote 6, Bloco H, Brasília - DF, 70070-912, priscilasena@ibict.br

(2) Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), SAUS Quadra 5 - Lote 6, Bloco H, Brasília - DF, 70070-912, tatyanesilva@ibict.br

(3) Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), SAUS Quadra 5 - Lote 6, Bloco H, Brasília - DF, 70070-912, julianasousa@ibict.br

(4) Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), SAUS Quadra 5 - Lote 6, Bloco H, Brasília - DF, 70070-912, washingtonsegundo@ibict.br

(5) Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), SAUS Quadra 5 - Lote 6, Bloco H, Brasília - DF, 70070-912, bianca@ibict.br

Resumo

A Ciência Aberta abarca práticas voltadas para a transparência da pesquisa científica, promovendo o detalhamento dos processos de elaboração de metodologias e gestão de dados científicos. Posto isso, o objetivo de relatar o processo de elaboração de uma proposta de critérios de qualificação de repositórios de dados e de publicações, torna-se relevante e pertinente para a comunidade acadêmica e industrial que trabalham com dados no Brasil. Justifica-se por se tratar de um dos 11 Marcos do Compromisso 8 - Construir uma proposta de modelo de avaliação que fomente a Ciência Aberta (Marco 2 - Proposição de critérios de Qualificação de Repositórios de Dados, de Repositórios de Publicações), no 5º Plano de Ação Nacional, na Parceria para Governo Aberto (OGP). Utilizou-se de um relato de experiência, pautado no registro documental-analítico, o que permite a caracterização como uma pesquisa documental e descritiva, de abordagem qualitativa. Como resultado se descreveram as três atividades estabelecidas para o desenvolvimento da proposta. O grupo de trabalho envolvido espera produzir um conjunto de recomendações normativas para auxiliar, principalmente, gestores de repositórios a fazerem uma avaliação da qualidade e confiabilidade dos repositórios de dados e/ou de publicações.

Palavras-chave: Ciência Aberta; Qualificação de repositórios; Repositório de Dados; Repositórios de publicações.

Abstract

Open Science embraces practices aimed at the transparency of scientific research, promoting the detailing of the processes of elaboration of methodologies and management of scientific data. Therefore, the objective of reporting on the process of developing a proposal of qualification criteria for data and publication repositories becomes relevant and pertinent to the academic and industrial community that works with data in Brazil. It is justified because it is one of the 11 Milestones of Commitment 8 - Build a proposal for an evaluation model that fosters Open Science (Milestone 2 - Propose Qualification Criteria for Data and Publication Repositories), in the 5th National Action Plan, in the Open Government Partnership (OGP). An experience report was used, based on a documental-analytical register, which allows characterizing this work as a documental and descriptive research, with a qualitative approach. As a result, the three activities established for the development of the proposal were described. The working group involved hopes to produce a set of normative recommendations to help, mainly, repository managers to make an evaluation of the quality and reliability of data and/or publication repositories.

Keywords: Open Science; Repository Qualification; Data Repository; Publication Repositories.

1 Introdução

Conforme o Digital Repositories JISC Briefing Paper (2005), um repositório digital é onde conteúdos digitais e recursos estão armazenados, e podem ser pesquisados e recuperados para uso futuro. Desse modo, suporta mecanismos de importação, exportação, identificação, armazenamento e recuperação de recursos digitais.

Os repositórios digitais constam definidos no Tesauro Brasileiro de Ciência da Informação, como dispositivos para "administrar, armazenar e preservar conteúdos informacionais em formato eletrônico", e que podem ser norteados por um assunto (repositórios temáticos) ou pela produção científica de uma instituição (repositórios institucionais) (PINHEIRO; FERREZ, 2014, p. 195).

Compreende-se os repositórios como ferramentas que registram o desenvolvimento da pesquisa por meio dos dados coletados, e ainda podem auxiliar a metodologia científica, em casos de experimentos sem êxito, cujos dados também podem ser registrados (SAYÃO; SALES, 2016).

Evidencia-se nas definições de repositórios digitais, alinhamento com as abordagens de Ciência Aberta, uma vez que, como um "movimento dos movimentos" (ALBAGLI, 2019), ela abarca práticas voltadas para a transparência da pesquisa científica, promovendo o detalhamento dos processos de elaboração de metodologias e gestão de dados científicos, para que estes sejam distribuídos, reutilizados e estejam acessíveis sem custos a todos os níveis da sociedade (SILVA; SILVEIRA, 2019).

Os repositórios de dados possibilitam o armazenamento e a gestão de dados vislumbrando o aprimoramento da recuperação, o que aumenta as potencialidades de reuso dos dados entre os pesquisadores. Deste modo, colabora para otimizar os processos da investigação científica, resultando no avanço da ciência (MONTEIRO, 2021).

Os repositórios de publicações embora sem conceito específico, alicerçam-se nas definições de repositórios institucionais. De acordo com Crow (2002), os repositórios institucionais são uma coleção digital que captura e preserva a produção intelectual da comunidade acadêmica. De forma que devem estar focados no armazenamento e promoção do acesso aberto para as publicações revisadas pelos pares (HANARD, 2001). A partir do posicionamento dos autores citados e para compreensão do trabalho apresentado, considera-se que nos repositórios de publicações os conjuntos de dados de pesquisa não fazem parte de seu escopo.

Posto isso, relatar o processo de elaboração de uma proposta de critérios de qualificação de repositórios de dados e de publicações é o objetivo delimitador deste trabalho.

Este objetivo é justificado por se tratar também de um dos 11 Marcos do Compromisso 8 - Construir uma proposta de modelo de avaliação que fomente a Ciência Aberta (Marco 2 - Proposição de critérios de Qualificação de Repositórios de Dados, de Repositórios de Publicações), no 5º Plano de Ação Nacional, na Parceria para Governo Aberto (do inglês *Open Government Partnership* - OGP) (TRANSPARÊNCIA EM CIÊNCIA…, 2022). Logo, torna-se relevante e pertinente para a comunidade acadêmica e industrial que trabalham com dados no Brasil, conforme visa o Workshop de Informação, Dados e Tecnologia (WIDaT), especificamente no subtópico "Ciência Aberta" do tópico "Dados".

2 Procedimentos Metodológicos

Relatar um processo constitui-se em uma ferramenta valiosa para a compreensão da diversidade intrínseca à ciência contemporânea. O Relato de Experiência, de acordo com Daltro e Faria (2019), pode ser empregado como um instrumento político/social, devido ao registro documental-analítico realizado. O que permite caracterizar este trabalho como uma pesquisa documental e descritiva, de abordagem qualitativa.

Para tanto, o processo aqui detalhado refere-se às perspectivas de oito instituições diferentes: Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Comissão Nacional de Energia Nuclear (CNEN) Associação Brasileira de Editores Científicos (Abec), *Scientific Electronic Library Online* (SciELO), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação Oswaldo Cruz (Fiocruz), Empresa Brasileira de Pesquisa Agropecuária (Embrapa) e Universidade Estadual Paulista"Júlio de Mesquita Filho" (Unesp)

O Ibict como órgão/entidade responsável por coordenar a execução do Marco 2 - Proposição de critérios de Qualificação de

Repositórios de Dados, de Repositórios de Publicações, optou por utilizar a Plataforma Google Meet[1] para realizar as reuniões online com as demais instituições.

3 Resultados Preliminares

Postas as definições utilizadas para a compreensão do grupo de trabalho em relação à repositório de dados e a repositório de publicações, foram estabelecidas as seguintes atividades para alcançar a concretização esperada para o Marco 2, a saber:

Atividade 1: Verificação de um conjunto inicial de critérios (exemplo: existência de identificadores persistentes, política de preservação). Essa atividade teve como ponto de partida as diretrizes OpenAIRE para repositórios de dados[2] e o quadro para boas práticas em repositórios da comunidade COAR[3] além do conhecimento e das experiências práticas dos integrantes do grupo. O Quadro 1 contém os indicadores levantados pelos integrantes do grupo e pela literatura.

Quadro 1 - Indicadores preliminares do Marco 2 do Compromisso 8 do 5º Plano de Ação brasileiro na Parceria para Governo Aberto

Nº	Indicadores
1	Número de acessos ao repositório
2	Número de downloads no repositório
3	Percentual de identificadores perenes existentes no repositório
4	Qualidade dos metadados
5	Percentual de registros que permitem acesso ao texto completo ou aos conjuntos de dados
6	Existência de instrumentos de gestão; política
7	Serviços ofertados pelo repositório, como ferramentas de análise de dados, modelagem, estatísticas de uso dos dados, consultoria para elaboração de PGD, serviços de referência para encontrar outros dados, mecanismos que auxiliam a citação
8	Interoperabilidade
9	Visibilidade aos dados
10	Recompensa do pesquisador por meio de estatística de citação
11	Acompanha os avanços tecnológicos
12	Atende as demandas do pesquisador

Fonte: Dados da Pesquisa (2022)

Atividade 2: Pesquisa sobre a literatura existente (levantamento bibliográfico e de grupos existentes no mesmo tema); considerando também as tecnologias envolvidas (Exemplo: *DSpace* 7 e *Dataverse*) e pesquisa de critérios já levantados e aplicados sobre o ARCA Fiocruz (ISO-16363 e critérios *Trustworthy Repositories Audit and Certification: Criteria and Checklist* - TRAC / TRUST / FAIR para Repositórios de Dados).

Atividade 3: Utilização da matriz do *Resource Description and Access* (RDA)[4] para criação de critérios destinados a repositórios de dados e para repositórios em geral. A primeira fase compreendeu a tradução das sugestões da matriz do inglês para o português. Seguindo para a separação das sugestões em colunas: "Quem", "O que", "Para quê", "Tipo de critério", "Aplicável somente a repositórios de dados?". A partir dessa separação, os integrantes apresentaram sugestões de critérios que atendesse a demanda de cada perfil e sinalizaram se aquele critério poderia ser aplicado para repositórios em geral.
Ainda com base na matriz de critérios da RDA, o grupo identificou e definiu os papéis que cada perfil profissional tem dentro do cenário que compõe o universo dos repositórios de dados e de publicação.

[1] Mais informações em: https://meet.google.com/.

[2] https://livroaberto.ibict.br/bitstream/123456789/1086/2/Diretrizes%20OpenAIRE%20para%20repositórios%20de%20dados.pdf

[3] https://livroaberto.ibict.br/bitstream/123456789/1089/2/Quadro%20para%20Boas%20Pr%c3%a1ticas%20em%20Reposit%c3%b3rios%20da%20Comunidade%20COAR.docx.pdf

[4] Ver mais em: https://miro.com/app/board/uXjVOE-p7AI=/.

4 Considerações Finais

Ao considerar as definições de repositório digital adotadas no decorrer do texto e os objetivos da Ciência Aberta, que tem como princípio fundamental o acesso livre aos conjuntos de dados de pesquisa e aos resultados da pesquisas financiadas com recursos públicos, evidencia-se que a implementação e a gestão de serviços que garantem a exibição, recuperação, exportação, identificação, preservação, uso e reuso de recursos digitais vão exigir iniciativas que apoiem a disseminação de boas práticas para auxiliar os repositórios a alcançarem um nível eficaz no que tange a interoperabilidade.

Nessa direção, o grupo espera produzir um conjunto de recomendações normativas para auxiliar, principalmente, gestores de repositórios a fazerem uma avaliação da qualidade e confiabilidade dos repositórios de dados e/ou de publicações. De igual modo, almeja-se contribuir para a criação de novos repositórios estruturados em consonância com a Ciência Aberta.

Dessa maneira, vislumbra-se que a publicação, disseminação e adoção das recomendações possam proporcionar maior adequação dos repositórios brasileiros às práticas de compartilhamento e interoperabilidade disseminadas por iniciativas internacionais.

Por fim, conclui-se que o objetivo determinado para este trabalho se concretizou, uma vez que se obteve o relato almejado do processo de elaboração de uma proposta de critérios de qualificação de repositórios de dados e de publicações. Proposta essa basilar para a efetivação do Compromisso 8 - Construir uma proposta de modelo de avaliação que fomente a Ciência Aberta. Isso porque, as publicações e dados científicos, bem como seus processos de gestão são imprescindíveis para o impulsionamento e sustentabilidade da Ciência Aberta.

Agradecimentos

À Controladoria-Geral da União por toda a gestão e suporte para a concretização do Compromisso 8 do 5º Plano de Ação Brasileiro.

À todas as pessoas e instituições participantes do grupo de trabalho referente ao Marco 2 do Compromisso.

À Financiadora de Estudos e Projetos (Finep) pelo subsídio financeiro concedido.

Referências

ALBAGLI, Sarita. O que é Ciência Aberta e qual o papel das agências de fomento diante deste fenômeno? In: Encontro Capes de Ciência Aberta. Tema: direitos de propriedade intelectual e políticas institucionais. Brasília, [s. n.], 2019. Disponível em: http://capes.gov.br/conteudo/2-encontro-capes-de-ciencia-aberta/.

CROW, Raym *et al.* The case for institutional repositories: A SPARC position paper. ARL Bimonthly Report, n. 223, 2002. Disponível em: https://ils.unc.edu/courses/2014_fall/inls690_109/Readings/Crow2002-CaseforInstitutionalRepositoriesSPARCPaper.pdf

DALTRO, Mônica Ramos; FARIA, Anna Amélia de. Relato de experiência: Uma narrativa científica na pós-modernidade. **Estudos e pesquisas em psicologia**, v. 19, n. 1, p. 223-237, 2019. Disponível em: https://www.e-publicacoes.uerj.br/index.php/revispsi/article/view/43015.

DIGITAL REPOSITORIES JISC. **Digital repositories**: helping universities and colleges. 2005. Disponível em: https://www.yumpu.com/en/document/view/15608201/digital-repositories-jisc.

HARNAD, Stevan. Research access, impact and assessment. **Times higer education supplement**, n. 18. maio 2001. Disponível em: https://eprints.soton.ac.uk/255950/1/thes1.html

MONTEIRO, Elizabete Cristina de Souza de Aguiar. **Operacionalização de repositórios de dados:** uma análise sobre as perspectivas e atitudes dos pesquisadores nas questões de autoria e licença. Tese (Doutorado em Ciência da Informação).

Marília/SP: Universidade Estadual Paulista (Unesp), Faculdade de Filosofia e Ciências, 2021. 268 f. Disponível em: https://repositorio.unesp.br/handle/11449/214671.

PINHEIRO, Lena Vania Ribeiro; FERREZ, Helena Dodd. **Tesauro Brasileiro de Ciência da Informação**. Rio de Janeiro; Brasília: Instituto Brasileiro de Informação em Ciência e Tecnologia, 2014. Disponível em: http://sitehistorico.ibict.br/publicacoes-e-institucionais/tesauro-brasileiro-de-ciencia-da-informacao-1.

SAYÃO, Luis Fernando; SALES, Luana Farias. Algumas considerações sobre os repositórios digitais de dados de pesquisa. **Informação & Informação**, v. 21, n. 2, p. 90-115, 2016. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/viewFile/27939/20122.

SILVA, Fabiano Couto Corrêa da; SILVEIRA, Lúcia da. O ecossistema da Ciência Aberta. **Transinformação**, v. 31, 2019. Disponível em: https://www.scielo.br/j/tinf/a/dJ89vRg94Qxtf6Y7M49Hztr/?lang=pt.

TRANSPARÊNCIA EM CIÊNCIA PARA O AVANÇO DA CIÊNCIA ABERTA - 5º PLANO DE AÇÃO OGP Brasil. 2022. Disponível em: https://wiki.rnp.br/x/So-QCQ.

REPOSITÓRIOS DE DADOS DE PESQUISA: ANÁLISE À LUZ DOS PRINCÍPIOS FAIR

RESEARCH DATA REPOSITORY: ANALYSIS REGARDING FAIR PRINCIPLES

Letícia Guarany Bonetti[1], Ana Carolina Simionato Arakaki[2]
(1) Universidade Federal de São Carlos (UFSCar), E-mail: leticia.bonetti@estudante.ufscar.br.
(2) Universidade Federal de São Carlos (UFSCar), E-mail: acsimionato@ufscar.br.

Resumo
Os princípios FAIR, essenciais para a maximização do potencial dos dados para a ciência, buscam aumentar a encontrabilidade, acessibilidade, interoperabilidade e reutilização dos dados, enfatizando o processamento automático por máquinas. O objetivo deste trabalho é, portanto, avaliar o nível de conformidade dos dados de pesquisa depositados nos repositórios institucionais do Estado de São Paulo quanto aos princípios FAIR. Trata-se de uma pesquisa exploratória e descritiva, com uma abordagem quantitativa. Para a verificação da aderência aos princípios FAIR foi utilizada uma ferramenta auxiliar, a *F-UJI Automated FAIR Data Assessment Tool*. Os níveis de aderência aos princípios FAIR dos conjuntos de dados analisados ainda se encontram baixos. Os dados em maior conformidade obtiveram pontuação de 50% de aderência. Já a menor pontuação geral foi 18%. Percebe-se que os repositórios têm pontos fracos em comum, principalmente quanto à reutilização e à interoperabilidade. Conclui-se que os resultados seguem uma tendência internacional, e todos os repositórios precisam buscar aperfeiçoar algumas funcionalidades.

Palavras-chave: Dado de pesquisa; Repositório de dados; Princípios FAIR.

Abstract
The FAIR principles, essential for maximizing the potential of data for science, seek to increase the findability, accessibility, interoperability and reuse of data, associated with research that is machine readable. This work aims to assess the FAIRness of research data deposited in institutional repositories in the State of São Paulo. It is an exploratory and descriptive research, with a quantitative approach. To verify compliance with the FAIR principles, an auxiliary tool, the F-UJI Automated FAIR Data Assessment Tool, was used. The levels of FAIRness of the analyzed datasets are low. The most compliant data obtained a 50% adherence score. The lowest overall score was 18%. It is noticed that the repositories have weaknesses in common, mainly in terms of reuse and interoperability. In conclusion, the results follow an international trend, and all the repositories need to improve some functionalities.

Keywords: Research data; Data repository; FAIR Principles.

1 Introdução

Com os avanços das tecnologias da informação e da comunicação (TICs), há também um aumento exponencial do volume de dados produzidos pelos pesquisadores, os chamados dados de pesquisa. Esses dados, quando compartilhados, trazem vários benefícios para a ciência, como uma maior transparência, reprodutibilidade, economia de recursos e agilização do ciclo científico.

De acordo com Borgman, Scharnhorst e Golshan (2019), muitas partes interessadas já estão envolvidas nas infraestruturas associadas aos dados de pesquisa, em busca da maximização do seu potencial para os avanços científicos. Algumas delas são: pesquisadores, agências de financiamento, universidades, formuladores de políticas de pesquisa e os usuários desses dados.

Uma das opções de infraestrutura amplamente utilizadas para a gestão dos dados, principalmente quando se fala das instituições de ensino superior nacionais, são os repositórios. Mas uma gestão adequada deve ir além do seu armazenamento e acesso, estando ligada às boas práticas internacionalmente adotadas, como é o caso dos princípios FAIR (SALES *et al.,* 2020).

Os princípios FAIR são um acrônimo para "*findable*", "*accessible*", "*interoperable*" e "*reusable*". Os quatro princípios foram estabelecidos como resultado da conferência internacional *'Jointly designing the data FAIRPORT*', de 2014. A conferência reuniu especialistas de diversos países e áreas para discutir uma infraestrutura global para publicação, descoberta, compartilhamento e reutilização de dados.

O repositório é essencial no ecossistema de dados FAIR (HODSON *et al*., 2018), que buscam enfatizar o aprimoramento da capacidade das máquinas de encontrar e processar os dados de forma automática (WILKINSON *et al.,* 2016). Isso se torna fundamental quando se fala na quantidade massiva de dados produzidos, que vão muito além do processamento humano. Sendo assim, para aumentar a encontrabilidade, a acessibilidade, a interoperabilidade e a reutilização dos dados de pesquisa, e melhorar o ecossistema desses dados, é preciso falar da adoção dos princípios FAIR.

2 Objetivos

Com isso, o objetivo desse trabalho é avaliar o nível de conformidade dos dados de pesquisa depositados nos repositórios institucionais do Estado de São Paulo quanto aos princípios FAIR, em foco a UFSCar, a Unicamp e a USP.

3 Procedimentos Metodológicos

A pesquisa caracterizou-se como exploratória e descritiva. Por meio de uma abordagem quantitativa buscou avaliar os conjuntos de dados de pesquisa depositados nos repositórios institucionais sob a ótica dos princípios FAIR.

A amostra foi definida a partir do metabuscador de dados de pesquisa da Fundação de Amparo à Pesquisa do Estado de São Paulo, que reúne repositórios de dados de pesquisa de instituições da região. Estavam mapeados nove repositórios em setembro de 2022. Entretanto, por se tratar de um trabalho derivado da dissertação da autora, ainda em andamento, foram considerados apenas os três repositórios analisados até o momento, buscando sistematizar achados e fazer comparações.

Para a verificação da aderência dos dados aos princípios FAIR foi utilizada uma ferramenta auxiliar, a *F-UJI Automated FAIR Data Assessment Tool*[1]. É um serviço *web* para avaliar programaticamente o nível de aderência dos dados. Ela foi desenvolvida pela *Fostering Fair Data Practices in Europe* – FAIRsFAIR, que desenvolve padrões globais para a certificação FAIR dos dados, sendo um projeto de extrema relevância no cenário internacional. A ferramenta é baseada em *Representational State Transfer* (REST) para a avaliação automatizada dos dados. Para isso foi preciso inserir na ferramenta o identificador do conjunto de dados que se deseja avaliar.

Neste trabalho o foco é analisar a média de notas que os conjuntos de dados alcançaram quanto aos quatro princípios FAIR, de forma geral; bem como sua pontuação em cada um dos princípios, de forma individual. A pontuação geral varia de 0 a 100%, e apresenta uma visão mais ampla dos repositórios da amostra. Já a avaliação individual apresenta um panorama dos pontos fortes e fracos, indicando o nível da aderência quanto a cada um dos quatro princípios, sendo a escala da ferramenta F-UJI: inicial, moderado ou avançado. Todos esses dados (percentual e nível de aderência) são entregues automaticamente pela ferramenta F-UJI, com métricas baseadas nos indicadores propostos pelo *RDA FAIR Data Maturity Model Working Group*[2] e no *WDS/RDA Assessment of Data Fitness for use checklist*[3]. As métricas, os métodos, as escalas e o código podem ser consultados em detalhes no *site* da ferramenta F-UJI[4].

Dessa forma, todos os 199 conjuntos de dados que estavam depositados nos três repositórios foram individualmente e automaticamente avaliados pela ferramenta auxiliar. Após a avaliação, os dados retornados foram compilados em planilhas para análises e comparações, o que possibilitou apontar pontos fortes e fracos de cada repositório baseado nos números e níveis entregues pela ferramenta F-UJI. Sendo assim, o escopo deste estudo limita-se aos resultados encontrados a partir da análise citada, possibilitando uma visão geral sobre o cenário regional de aderência aos princípios FAIR e trazendo um *feedback* para as instituições avaliadas.

[1] Disponível em: https://f-uji.net.

[2] Disponível em: https://zenodo.org/record/3909563#.Y2AKly35RhA

[3] Disponível em: https://doi.org/10.15497/RDA00034.

[4] Disponível em: https://www.f-uji.net/index.php?action=about

4 Resultados

4.1 Repositório da UFSCar

O repositório da Universidade Federal de São Carlos (UFSCar) continha um total de 15 conjuntos de dados no momento da coleta. Com relação à aderência aos princípios, no geral, a maior nota alcançada foi 35%, e a menor foi 27%. Apenas três conjuntos de dados conseguiram alcançar a nota mais alta e a média de pontuações foi 30,8%.

Esses dados iniciais revelam uma falta de aderência geral aos princípios FAIR, com uma certa padronização entre as pontuações alcançadas pelos 15 conjuntos de dados. O depósito é feito por autoarquivamento, guiado por um manual institucional, o que pode explicar os resultados similares, uma vez que a representação está intimamente ligada com os resultados de aderência.

Na análise individual é possível verificar que os dois princípios com maior dificuldade para aderência foram o interoperável (maior nota sendo 1/4) e reutilizável (maior nota sendo 4/10). Dunning, Smaele e Böhmer (2017) corroboram esse resultado ao afirmar que, em avaliação feita em 38 repositórios da Holanda, “interoperável” e “reutilizável” foram os princípios mais difíceis de aderir. 38% dos repositórios não possuíam metadados ricos e apenas 41% atribuíam uma licença clara.

No princípio **encontrável,** a maior nota foi 4/7, sendo que a maioria recebeu a pontuação três. Os metadados não incluíam o identificador dos dados que descreviam, e não foi possível encontrar um identificar persistente ou um *Uniform Resource Locator* (URL) que indicava a localização do conteúdo de dados para o *download*.

Os identificadores persistentes são de suma importância na encontrabilidade dos dados, dando suporte para que as citações e reutilizações sejam rastreáveis. Mas de acordo com a ferramenta F-UJI, apesar de todos os dados possuírem um identificador único e global, apenas alguns possuíam um identificador persistente. Juty *et al.* (2020) afirmam que existem quatro tipos comuns de identificadores FAIR: *Digital Object Identifier* (DOI), *Archival Resource Key* (ARK), *Identifiers.org* e *Persistent Uniform Resource Locator* (PURL), que não foram encontrados para a maioria dos dados da UFSCar.

Em compensação, de acordo com a ferramenta auxiliar, os conjuntos de dados possuíam um nível moderado de elementos de metadados para sua descrição. O uso de metadados ricos é essencial para que eles sejam encontrados na *web* e posteriormente interpretados por pesquisadores e máquinas.

Em **acessível,** todos os conjuntos de dados obtiveram nota 1,5/3. O princípio está relacionado com protocolos bem definidos e universalmente implementáveis. Mas os metadados não incluíam um *link* resolvível para os dados com base em protocolos de comunicação da *web* padronizados; e apesar de ter encontrado informações sobre direitos de acesso nos metadados, elas não eram legíveis por máquina, indicando um nível inicial a ser aprimorado pela UFSCar.

Em **interoperável**, muitos conjuntos de dados obtiveram nota zero. Silva e Santarem Segundo (2021) afirmam que este princípio está ligado com a atribuição de metadados interligados e a capacidade de sistemas se comunicarem. Segundo Henning *et al.* (2018), é o princípio FAIR mais desafiador.

Quando se fala em interoperabilidade, é importante citar a necessidade do uso de metadados estruturados como *JavaScript Object Notation* (JSON) e *Resource Description Framework* (RDF) incorporados no código XHTML/HTML, o que não foi identificado pela ferramenta. Entretanto, o fato de existirem recursos relacionados nos metadados (“*isRelatedTo”*) é um ponto positivo. Mas essa funcionalidade não foi identificada para todos os conjuntos de dados, e por isso recomenda-se que a instituição busque incentivá-la.

Já quanto à **reutilização**, há um nível inicial de especificação do conteúdo dos dados nos metadados. Essas informações são essenciais para a contextualização dos dados de pesquisa, e por isso é importante que a instituição aprimore a descrição com metadados como tamanho e tipo do arquivo em formato legível por máquina. Entretanto, quanto às informações sobre a proveniência dos dados, o nível foi considerado moderado, extraindo elementos como criador, data de publicação, editor e contribuidor.

Outro aspecto importante é a declaração da licença sob as quais os dados podem ser reutilizados. A ferramenta não conseguiu identificar uma licença declarada em um

elemento de metadados apropriado para a maioria dos dados de pesquisa. Nem uma licença registrada no *Software Package Data Exchange* (SPDX), uma lista de licenças usadas em *software*s, dados, *hardwares* ou documentação abertos ou colaborativos.

Apesar de existir a indicação da licença *Creative Commons* nas páginas dos itens, apenas dois conjuntos de dados foram validados, apresentando a mesma licença: http://creativecommons.org/publicdomain/zero/1.0/. Recomenda-se que a instituição busque padronizar essa indicação, para que todos os dados tenham essa informação bem explícita para humanos e máquinas. Um ponto positivo é que os metadados seguem um padrão recomendado pela comunidade, o *Dublin Core*.

Sendo assim, os conjuntos de dados do repositório da UFSCar foram avaliados pela ferramenta, no geral, com um nível inicial de FAIR. Recomenda-se a observância dos aspectos ligados à interoperabilidade e à reutilização, como a declaração padronizada nos metadados da licença sob a qual os dados podem ser reutilizados. Além disso, a adoção de identificadores persistentes como o DOI, e a ampliação de metadados indicados para preenchimento, garantindo uma descrição mais rica dos dados de pesquisa, que demandam um alto grau de contextualização.

4.2 Repositório da Unicamp

O repositório da Universidade Estadual de Campinas (Unicamp) continha um total de 67 conjuntos de dados no momento da coleta. Com relação à aderência geral, a maior nota obtida foi 50%, e a menor 45%, demonstrando novamente uma padronização para conjuntos de dados depositados no mesmo repositório. A média de pontuações foi 49,5%. Percebe-se uma melhoria de resultado quando comparado à UFSCar.

A Unicamp possui uma Comissão de Gestão de Dados de Pesquisa (CGDP), com a incumbência de promover a política institucional de dados de pesquisa, propondo ações segundo as boas práticas nacionais e internacionais. A instauração da Comissão pode estar relacionada com os melhores resultados encontrados.

No princípio **encontrável**, todos os conjuntos de dados obtiveram pontuação 6/7. A todos os conjuntos de dados foram atribuídos um DOI. Isso garante uma maior encontrabilidade, além de permitir a citação persistente. Os metadados, assim como no caso da UFSCar, não incluíam o identificador dos dados que descreviam, e não foi possível encontrar um PID ou URL que indicava a localização do conteúdo de dados para o *download.* Isso demonstra um ponto comum a ser aperfeiçoado por ambas as instituições.

Em compensação, todos os dados de pesquisa depositados possuíam um nível avançado de metadados com elementos descritivos essenciais, como criador, título, data de publicação, resumo e palavras-chave. O uso de metadados ricos é essencial neste princípio (HODSON *et al*., 2018).

Quanto à **acessibilidade**, a pontuação foi 1/3 para todos os dados de pesquisa avaliados. Foi o único princípio em que a Unicamp obteve notas menores que as da UFSCar. A ferramenta apontou que não houve a identificação do nível de acesso aos dados, mas ao consultar a página dos itens é possível encontrar os termos de acesso. Isso pode estar ligado ao fato de a informação não estar legível por máquina, já que os princípios buscam enfatizar a capacidade de processamento automático, e por isso é importante adotar soluções nesse sentido. Recomenda-se, portanto, que a instituição avalie como são indicados os termos de acesso em seus metadados.

Assim como no caso dos conjuntos de dados da UFSCar, os metadados da Unicamp não incluíam um *link* resolvível para os dados com base em protocolos de comunicação da *web* padronizados, outro ponto a ser aperfeiçoado.

Já em **interoperabilidade** as notas variaram entre os conjuntos de dados, com alguns alcançando nota dois e outros nota três de quatro. Diferente da UFSCar, a Unicamp obteve nível avançado para o uso de metadados estruturados incorporados no código XHTML/HTML. Ainda quanto a este princípio, a ferramenta foi capaz de localizar, em nível avançado, metadados com *links* entre os dados e suas entidades relacionadas ("*HasPart*"). Mas a maioria dos dados obtiveram um nível inicial quanto aos metadados utilizarem recursos semânticos como *namespaces*, que garantem que determinado conjunto de objetos tenha

nomes exclusivos para que possa ser facilmente identificado, de forma unívoca.

Já quanto à **reutilização**, todos os conjuntos de dados da Unicamp obtiveram nota igual a 2/10. Também foi identificado um nível inicial de especificação do conteúdo dos dados nos metadados. É importante que a instituição aprimore essa descrição. Em contrapartida, quanto às informações sobre a proveniência dos dados de pesquisa, o nível foi considerado moderado. A ferramenta F-UJI encontrou elementos como a data de publicação, a data de modificação, criador, editor e contribuidor.

Inclusive, fornecer informações sobre o versionamento dos dados, como é feito pela Unicamp, é uma boa prática incentivada pela *World Wide Web Consortium* (W3C), a principal organização de padronização da *Web*. É um ponto positivo que auxilia os usuários dos dados a determinar com qual versão estão trabalhando, e se a versão em questão é a mais atual. O versionamento explícito permite comparações e evita confusão com o uso.

A ferramenta F-UJI chegou a localizar os metadados de licença (*schema.org*), mas alegou que a representação está incorreta. É importante citar que a ferramenta faz a busca na lista SPDX, que procura identificar de forma confiável as licenças válidas. A sintaxe CC0, como é apresentada no repositório da Unicamp, não é compatível com a sintaxe fornecida pela lista. Apesar de ser legível por humanos, pode não ser pelas máquinas. A busca pela padronização é um aspecto importante quando se fala de dados FAIR, principalmente para a interoperabilidade.

Assim como no caso da UFSCar, os metadados seguem um padrão recomendado pela comunidade*: Dublin Core*. Sendo assim, os dados do repositório da Unicamp foram avaliados com um nível moderado de FAIR. Recomenda-se a observância dos aspectos com relação à acessibilidade e à reutilização, com as menores pontuações.

4.3 Repositório da USP

Já o repositório da Universidade de São Paulo (USP) continha um total de 117 conjuntos de dados no momento da coleta. Quanto à aderência aos quatro princípios, a maior nota foi 25% e a menor 18%. A média de notas foi 22,8%. Percebe-se que, quando comparados aos dados de pesquisa do repositório da UFSCar e da Unicamp, os conjuntos de dados da USP possuem um nível mais baixo de aderência ao FAIR.

No princípio **encontrável**, as pontuações variaram entre 2,5 e 3 de no máximo 7. Todos os dados avaliados possuíam um identificador único e global, mas nenhum deles foi validado como persistente. Também não incluíam o identificador dos dados que descreviam, seguindo a mesma tendência dos dois repositórios anteriores.

Quanto aos metadados com elementos descritivos essenciais, 40 conjuntos de dados obtiveram nível moderado, enquanto os outros 77 obtiveram nível inicial. Esses 40 apresentam alguns elementos adicionais como "*dc.coverage.temporal*", "*dc.publisher*" e "*dc.coverage.spatial*". Recomenda-se que a USP busque padronizar essa descrição mais rica, tornando mais elementos obrigatórios para a publicação dos dados de pesquisa, que demandam uma contextualização rica.

Quanto à **acessibilidade**, a pontuação foi 1/3 para todos os dados de pesquisa avaliados, assim como no caso da Unicamp. Novamente a ferramenta F-UJI apontou que não houve a identificação do nível de acesso ou das condições de acesso aos dados nos metadados, um aspecto importante de ser verificado pela instituição.

Assim como no caso dos conjuntos de dados da UFSCar e da Unicamp, os metadados não incluíam um *link* resolvível para os dados com base em protocolos de comunicação da *web* padronizados.

Já em **interoperabilidade,** todos os conjuntos de dados obtiveram pontuação igual a zero. Ou seja, não foram identificados metadados estruturados e nem metadados que incluíssem *links* entre os dados de pesquisa e suas entidades relacionadas. Entretanto, foi possível localizar *namespaces* extraídos dos metadados em nível inicial. É o princípio que mais demanda atenção, com o uso de padrões *web* como base das APIs e metadados com recursos semânticos (SILVA; SANTAREM SEGUNDO, 2021).

Por fim, com relação ao princípio **reutilizável**, os conjuntos de dados da USP obtiveram avaliação semelhante aos da UFSCar e Unicamp, com 113 conjunto de dados obtendo nota igual a 2/10. Percebe-se um alto nível de padronização entre as três

instituições quanto a este princípio FAIR, demonstrando que a reutilização é uma faceta de difícil aderência no contexto dos repositórios da amostra.

Assim como no caso das outras duas universidades, foi identificado um nível inicial de especificação do conteúdo dos dados nos metadados. Já quanto às informações sobre a proveniência dos dados, o nível foi considerado moderado, com a extração de elementos como o criador, o editor, e a data de publicação.

Não foi identificada uma licença declarada em um elemento de metadados apropriado, o que impossibilitou que a ferramenta determinasse as condições de reutilização dos dados de pesquisa. É importante refrisar esse ponto porque a reutilização exige que os dados sejam publicados com uma licença clara e acessível. As condições sob as quais os dados podem ser usados devem ser transparentes tanto para humanos quanto para máquinas (HODSON *et al.,* 2018).

Assim como nos outros casos, os metadados do repositório da USP seguem um padrão recomendado pela comunidade. Há a padronização do uso do *Dublin Core* entre as três instituições da amostra.

Sendo assim, os conjuntos de dados do repositório da USP foram avaliados com um nível inicial de FAIR. Recomenda-se a observância dos aspectos relacionados principalmente com a interoperabilidade, em que todos os dados obtiveram nota zero. O princípio está diretamente ligado com o processamento automático por máquina, o que é essencial para o FAIR. É preciso o uso de recursos semânticos, metadados estruturados, vocabulários e referências qualificadas a outros dados e metadados. Além disso, recomenda-se a adoção de identificadores persistentes como o DOI, para que os dados possam ser inequivocamente referenciados e citados; e a ampliação de metadados indicados para preenchimento, garantindo uma descrição mais rica dos dados de pesquisa.

5 Considerações Finais

Para que haja a maximização do potencial dos dados de pesquisa, é preciso mais do que o armazenamento e acesso. É preciso adotar as boas práticas que vêm sendo difundidas internacionalmente, como é o caso dos princípios FAIR. Eles buscam aumentar a encontrabilidade, acessibilidade, interoperabilidade e reutilização dos dados.

Nota-se que os níveis de aderência aos princípios FAIR dos dados analisados ainda se encontram baixos. Os dados em maior conformidade obtiveram aderência de 50%. Já a menor nota geral foi 18%. Percebe-se que os repositórios têm pontos fracos em comum, principalmente quanto à reutilização. Os dados precisam ser acompanhados de metadados ricos e uma licença clara e acessível, pontos que podem ser aperfeiçoados por todos os três repositórios.

A interoperabilidade também se mostrou como um princípio de difícil aderência para os repositórios da UFSCar e da USP, e é indicado o uso de referências qualificadas, com atenção aos padrões e vocabulários para que os dados sejam legíveis por máquina. Mas, como já visto, os princípios interoperável e reutilizável representam um desafio internacional para aderência.

Dados acessíveis por protocolos de comunicação padronizados e identificadores persistentes são outros dois aspectos a serem aperfeiçoados pelas instituições. Em compensação, os três repositórios possuíam metadados que especificavam o conteúdo dos dados e incluíam informações da proveniência dos dados, fundamentais para sua interpretação e posterior reutilização pelos pares. Além do uso de um esquema de metadados recomendado pela comunidade, mesmo que em estágios iniciais. Esses são pontos em comum a serem destacados.

Referências

BORGMAN, C. L; SCHARNHORST, A.; GOLSHAN, M. S. Digital data archives as knowledge infrastructures: mediating data sharing and reuse. **Journal of the Association for Information Science and Technology**, v. 70, n. 8, 2019. Disponível em:https://asistdl.onlinelibrary.wiley.com/doi/epdf/10.1002/asi.24172. Acesso em: 17 jun. 2022.

DUNNING, A.; SMAELE, M.; BÖHMER, J. **Are The Fair Data Principles Fair**? 2017. Disponível em: https://zenodo.org/record/321423. Acesso em: 27 jul. 2022.

HENNING, P. C. *et al*. Desmistificando os princípios fair: conceitos, métricas, tecnologias e aplicações inseridas no ecossistema dos dados fair. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**, v. 14, n. 3, p. 175-192, 2019. Disponível em: https://brapci.inf.br/index.php/res/v/150613. Acesso em: 20 abr. 2022.

HODSON, S. *et al.* **Turning FAIR into reality**. European Commission. Luxembourg: Publications Office of the European Union, 2018. Disponível em: https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf. Acesso em: 05 jan.2020.

JUTY, N. *et al*. Unique, persistent, resolvable: Identifiers as the foundation of FAIR. **Data Intelligence**, v. 2, n. 1-2, p. 30-39, 2020. Disponível em: https://direct.mit.edu/dint/article/2/1-2/30/9992/Unique-Persistent-Resolvable-Identifiers-as-the. Acesso em: 30 jun. 2022.

SALES, L. *et al*. GO FAIR Brazil: A Challenge for Brazilian Data Science. **Data Intelligence**, v. 2, n. 1–2, p. 238–245, 2020. Disponível em: https://direct.mit.edu/dint/article/2/1-2/238-245/10004. Acesso em: 27 jul. 2022.

SILVA, L. C.; SANTAREM SEGUNDO, J. E. Princípios FAIR e Linked Data: publicação de cadernos abertos de pesquisa. *In:* SALES, L. F.; VEIGA, V. S. O.; HENNING, P.; SAYÃO, L. F. (org). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: IBICT, 2021. Disponível em: http://ridi.ibict.br/handle/123456789/1182. Acesso em: 05 set. 2022.

WILKINSON, M. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. **Sci Data**, n. 3, 2016. Disponível em:https://www.nature.com/articles/sdata201618. Acesso em: 20 jan. 2022.

REPRESENTAÇÃO DA INFORMAÇÃO EM MUSEUS: UMA REVISÃO SISTEMÁTICA EM BASES DE DADOS BRASILEIRAS

INFORMATION REPRESENTATION IN MUSEUMS: A SYSTEMATIC REVIEW IN BRAZILIAN DATABASES

Bruna Stefane de Freitas[1], Kris Ellen das Neves Teixeira[2], Luísa Vernersbach Varejão[3], Silvana Pires Rocha Nogueira[4], Daniela Lucas da Silva Lemos[5], Dalton Lopes Martins[6].

(1) Universidade Federal do Espírito Santo, Vitória, ES-Brasil,bruna.s.freitas@edu.ufes.br
(2) Universidade Federal do Espírito Santo, Vitória, ES-Brasil, kris.teixeira@edu.ufes.br
(3) Universidade Federal do Espírito Santo, Vitória, ES-Brasil, luisa.varejao@edu.ufes.br
(4) Universidade Federal do Espírito Santo, Vitória, ES-Brasil, sileclel@gmail.com
(5) Universidade Federal do Espírito Santo, Vitória, ES-Brasil, daniela.l.silva@ufes.br
(6) Universidade de Brasília, Brasília, Brasil, daltonmartins@.unb.br

Resumo

O presente artigo apresenta os resultados da pesquisa realizada em bases de dados brasileiras por meio da compilação de dados sobre a representação da informação em museus. Para tal, buscou-se elaborar um protocolo de revisão sistemática de dados coletados com a temática de representação da informação, web semântica, padrão de metadados e a maneira que tem sido abordada no contexto dos museus, objetivando conhecer como a temática da representação da informação dentro do contexto dos museus brasileiros tem sido desenvolvida por pesquisadores dentro da literatura científica em periódicos, trabalhos acadêmicos apresentados em eventos técnicos científicos voltados para as áreas da Museologia e Ciência da Informação, e que estejam disponíveis nas bases de dados previamente selecionadas, incluindo a BDTD, a Brapci, Anais do ENANCIB e da SciELO Brasil. Metodologicamente, os dados dos trabalhos selecionados foram coletados e analisados segundo as premissas de análise de conteúdo de Bardin, método fundamental para a sistematização de conteúdo a partir de categorias analíticas. Como resultados, evidenciaram-se que as categorias com maior revocação são as que abordam os aspectos tradicionais no campo da "Museologia" e mesmo dentre estas, o número de artigos é relativamente baixo, porém, é possível notar uma ampliação nas publicações nos últimos anos acerca da temática, principalmente por conta da implementação de novas tecnologias para a disponibilização de acervos museológicos online, que necessitam de padronização nas informações dos objetos catalogados.

Palavras-chave: representação da informação; museus; revisão sistemática; bases de dados brasileiras; análise de conteúdo.

Abstract

This article presents the results of research carried out in Brazilian databases through the compilation of data on the representation of information in museums. To this end, we sought to develop a protocol for the systematic review of data collected with the theme of information representation, semantic web, metadata standard and the way it has been approached in the context of museums, aiming to know how the theme of information representation within the context of Brazilian museums has been developed by researchers within the scientific literature in journals, academic works presented at technical scientific events focused on the areas of Museology and Information Science, and which are available in previously selected databases, including BDTD , Brapci, Anais do ENANCIB and SciELO Brasil. Methodologically, the data from the selected works were collected and analyzed according to Bardin's content analysis premises, a fundamental method for the systematization of content based on analytical categories. As a result, it was shown that the categories with the greatest recall are those that address traditional aspects in the field of museology and even among these, the number of articles is relatively low, however, it is possible to notice an increase in publications in recent years about the theme, mainly due to the implementation of new technologies for the availability of museum collections online, which need standardization in the information of the cataloged objects.

Keywords: information representation; museums; systematic review; Brazilian databases; metadata; semantic web.

1 Introdução

A representação da informação é composta por elementos descritivos e temáticos que descrevem as características específicas de um objeto informacional. Historicamente, a representação esteve nas formações sociais dos povos e culturas do mundo, que vivenciaram e acompanharam os mais diversos contextos de transformação social ao longo do tempo. E falar sobre representação da informação remete a Aristóteles, um dos primeiros estudiosos a pensar o mundo e a categorizá-lo, com sua persistente busca de saber como o homem poderia caracterizar o conhecimento. Para representar a informação, torna-se necessário o uso de técnicas no campo da Organização e Representação da Informação e do Conhecimento, pois ajuda a modelar o domínio e sua representação, bem como a estruturar os sistemas de recuperação da informação (SRI), de modo que o registro do conhecimento recuperado seja útil e consistente (LANCASTER, 2004; INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2016). Nesse contexto, a representação da informação se torna um insumo indispensável para que a informação seja organizada, recuperada, que seja acessível o uso de seus registros, e, para tal, são utilizados processos e instrumentos que visem corroborar para essa finalidade (SILVA; LARA, 2021).

De acordo com o Ibram (2010; 2011), após um levantamento, constatou-se que há, no Brasil, mais de 3.000 museus com diversos tipos de acervo, dentre os quais 1.500 disponibilizaram informação sobre os tipos de instrumentos utilizados para organizar e representar seus objetos, fisicamente ou digitalmente; porém, percebe-se pouca evidência na literatura de dados concretos sobre o assunto.

O conhecimento de como os museus trabalham a representação da informação, das diversas mudanças tecnológicas que ocorreram na sociedade, e de como essas transformações influenciaram a forma como essas instituições desenvolvem suas atividades, desde atividades internas até a exposição de suas obras, é de extrema importância para o desenvolvimento contemporâneo do campo. Um exemplo desta contemporaneidade são as iniciativas como a Europeana (WINER; ROCHA, 2013) e os Acervos em Rede (BRASIL, 2021), ambos projetos que têm como intuito disponibilizar coleções digitais, incluindo a de museus, para toda a sociedade. Essa disponibilização implica na organização e representação de informações agregadas de itens de coleções de instituições interessadas (denominadas provedoras) em compartilhar seus dados em interface única no ambiente em rede da Internet. Portanto, este trabalho se justifica a partir da compreensão de iniciativas nacionais relacionadas a esta temática de modo a torná-la visível e mais acessível para pesquisadores interessados no desenvolvimento de estudos relacionados.

Posto isso, diante da reconhecida necessidade de se conhecer e disseminar as práticas de organização da informação no âmbito dos museus brasileiros, em especial, formula-se a questão de pesquisa em saber como as temáticas representação da informação, web semântica e estudos sobre metadados têm sido abordadas no contexto das pesquisas sobre museus na literatura das áreas da Museologia e da Ciência da Informação no Brasil?

2 Objetivos

A presente pesquisa tem como principal objetivo apresentar o resultado de uma revisão sistemática na literatura sobre como a temática da representação da informação dentro do contexto dos museus brasileiros tem sido desenvolvida por pesquisadores dentro da literatura científica em periódicos, trabalhos acadêmicos apresentados em eventos técnicos científicos voltados para as áreas da Museologia e Ciência da Informação (CI), dada a interdisciplinaridade entre ambas as áreas.

3 Procedimentos Metodológicos

Para a contemplação deste objetivo foram realizadas as seguintes ações metodológicas: delimitação da temática e termos de busca determinados na pesquisa; levantamento dos trabalhos nas bases de

dados selecionadas; avaliação dos artigos levantados e exclusão de trabalhos fora dos critérios de seleção; criação de categorias para a análise do conteúdo selecionado; e a partir das categorias, analisar o conteúdo dos trabalhos selecionados e discutir como cada temática tem sido trabalhada na literatura. A pesquisa ambientou-se em 4 (quatro) diferentes bases de dados selecionadas a partir dos critérios de nacionalidade e cobertura. Desse modo, foram selecionadas bases de dados brasileiras e que se destacam no campo da CI, a saber: a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), a Base de dados de Periódicos em Ciência da Informação (Brapci), Trabalhos do Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação (ENANCIB) dos anos de 1994 a 2019 e da *Scientific Electronic Library Online* (SciELO Brasil). A partir disso, foram selecionados os termos de busca baseados em vocabulários controlados e tesauros das áreas de Museologia e CI. Ao todo, foram designados 9 (nove) termos, os quais se fizeram representativos pela ação de leitura flutuante (BARDIN, 2016), em que há um primeiro contato com os documentos selecionados para a coleta de dados, a saber: Documentação museológica; Catalogação museológica; Catalogação documental; Gestão da informação em museus; Gestão da informação museológica; Tratamento documental em museus; Tratamento documental museal; Catalogação de objetos museais; Gestão de informações museológicas. Os trabalhos recuperados em cada base de dados foram submetidos a critérios de seleção para inclusão aplicando-se as regras oriundas da análise de conteúdo, conhecidas como regra da exaustividade e regra da representatividade ou exclusão. Após a busca, os trabalhos passaram por análise e foram utilizados alguns critérios de exclusão, previamente definidos,a saber: a) trabalhos duplicados; b) trabalhos incompletos; e c) Trabalhos cujos resumos passaram por uma leitura minuciosa e não apresentaram compatibilidade com a temática do estudo.

Desta forma, no total da busca, 32 trabalhos foram selecionados, sendo: 7 da BDTD, 8 da Brapci, 12 do Enancib, 3 da SciELO e outros 2 trabalhos que foram encaminhados pelo grupo de pesquisa, de acordo com o Apêndice B.

Os dados dos trabalhos selecionados foram coletados e analisados segundo as premissas do método análise de conteúdo de Bardin (2016), sendo realizada a codificação dos materiais, a qual trata de levantar uma linguagem que é comum entre o *corpus* selecionado e a análise do contexto. Posteriormente, o material foi enumerado e, por fim, conduzido o processo de categorização para a criação das 7 (sete) categorias analíticas determinadas pelo método Apêndice A, concebidas a partir do critério semântico oriundo do método, o qual se orienta pela temática envolvida na análise. A partir da criação das categorias, os documentos passaram por leitura completa e atenta, seleção e classificação de seu conteúdo.

4 Comportamento das Categorias Analíticas nas bases de dados

Na base da BDTD, as categorias de maior incidência foram: “Documentação museológica”; “Tratamento da informação (tradicional)”; “Tipologia documental” e “Representação descritiva”, as quais foram representadas em 7 (sete) trabalhos. Em sequência, a categoria “Representação temática” é contemplada em 6 (seis) 9 (nove) trabalhos, a categoria “Interdisciplinaridade” em 3 (três) e a categoria “Análise do conteúdo documental” é representada em 2 (dois) trabalhos.

A Brapci apresenta duas categorias que mais se repetiram, e foram contempladas em 8 trabalhos recuperados, são elas: “Tipologia documental” e “Representação descritiva”, seguidas de “Documentação museológica” e “Representação temática” com 5 (cinco) trabalhos; a categoria “Interdisciplinaridade” tem recorrência em 4 (quatro) trabalhos e as categorias “Análise do conteúdo documental” e “Tratamento da informação (tradicional)” ocorrem em 3 (três) dos 8 (oito) trabalhos selecionados.

Nos 12 trabalhos recuperados na base de dados do Enancib, entre os anos de 1994 a 2019, 11 trabalhos contemplaram a categoria

"Análise do conteúdo documental", seguidos de 9 (nove) trabalhos que contemplam a categoria "Documentação museológica". Quanto às categorias "Tratamento da informação (tradicional)" e "Representação descritiva", a recorrência é notada em 7 (sete) dos 12 trabalhos, e por fim a categoria "Tipologia documental" ocorre em, somente, 3 (três) trabalhos.

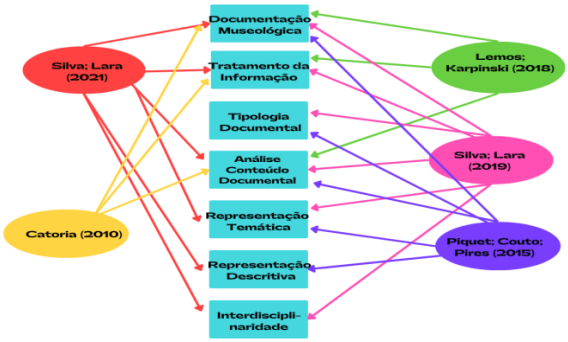
Em relação às outras bases de dados, a SciELO obteve a menor quantidade de trabalhos recuperados. Apenas 3 (três) artigos foram recuperados. Nesses trabalhos, as categorias "Análise do conteúdo documental" e "Representação descritiva" são contempladas em 2 (dois) trabalhos; as demais categorias ocorreram em 3 (três) dos 4 (quatro) trabalhos recuperados. Foram selecionados, à parte, 3 (três) trabalhos. Dentre eles, a categoria "Tipologia documental" não é contemplada. Em sequência, as categorias "Documentação museológica", "Tratamento da informação (tradicional)" e "Representação temática" têm recorrência em 2 trabalhos, por fim, as categorias "Análise do conteúdo documental", "Representação descritiva" e "Interdisciplinaridade" ocorreram em 1 trabalho.

O panorama das relações entre as bases de dados pesquisadas e as categorias analíticas pode ser visualizado no Apêndice B.

5 Revisão de literatura a partir das Categorias Analíticas

Por questão de delimitação de espaço no presente artigo, analisados na RSL, foram selecionados cinco (5) trabalhos considerados representativos em termos de princípios teórico-metodológicos presentes nas discussões aqui propostas. Os autores e as categorias encontradas nos textos estão representados na Figura 1:

Figura 1- Categorias versus Autoria



Fonte: os autores.

Silva e Lara (2021) apresentaram em seu trabalho uma análise de diretrizes de museus e propuseram a criação de um esquema de metadados para a catalogação de acervos museológicos, dissertando sobre a importância da política documental em museus como ferramenta essencial para a busca, guarda e recuperação das informações. Problematizam também a falta de conhecimentos relacionados a diretrizes de registros catalográficos e sua implicância no processo de catalogação dos acervos. Abordam a importância do uso de vocabulários controlados e dos registros dos procedimentos adotados na catalogação, para garantir a consistência e veracidade das informações. O trabalho de Catoria (2010) buscou analisar obras artísticas criadas pelo acervo de artes visuais e pelo Núcleo de Arte Contemporânea da Universidade Federal do Paraná (UFPR), dentro de um recorte de tempo. Ao longo do texto, a autora descreve a documentação como parte essencial do processo de musealização dos objetos. Em sequência, trabalha as definições de carga intrínseca e extrínseca dos objetos e sua importância no processo de tratamento da informação. Lemos e Karpinski (2018) realizaram um estudo sobre os trabalhos apresentados no ENANCIB dentro da modalidade oral, que relacionam a CI com a Museologia, especialmente nos aspectos de representação da informação. Os autores apresentaram definições sobre documentação museológica, chegando a conclusão que ela pode ser utilizada como fonte de informação. Também trabalharam as definições de carga intrínseca e extrínseca.

No trabalho de Piquet, Couto e Pires (2015) realizou-se um estudo de caso sobre a

implementação e desenvolvimento no uso da base de dados Personal Home Library (PHL), sua customização e as mudanças que ocorreram na troca de *software*. Os autores trabalharam as definições de documentação museológica e apresentaram as especificidades relacionadas ao tratamento dos objetos do acervo, que precisam de adaptações nos campos descritores do PHL, para que a descrição dos objetos seja feita de maneira mais completa.

Por fim, Silva e Lara (2019) apresentam um levantamento sobre as características de alguns museus de arte brasileiros. Em sequência, fazem uma análise de metadados propostos por diretrizes internacionais. As autoras apresentam os tipos de registros sobre objetos museológicos recomendados pelo IBRAM, e como as instituições realizam seus registros e os *softwares* que são utilizados. As autoras realizam a junção a análise no modo como os museus estão realizam o registro de seus acervos e sequencialmente faz uma comparação entre recomendações internacionais de padronização de metadados de artes visuais, demonstrado o potencial destes, como materiais de orientação para a criação de práticas de padronização para as instituições nacionais. As autoras apresentam um modelo de metadados bastante completo e de fácil utilização, para um registro mais completo do acervo.

6 Considerações Finais

O protocolo de revisão sistemática foi um passo importante para a definição de categorias e realização de análises dos artigos selecionados de acordo com a temática investigada. Desse modo, o problema de pesquisa e seu objetivo foram atingidos a partir do conhecimento sobre como a temática da representação da informação tem sido trabalhada dentro do contexto dos museus brasileiros. Para isso, foram elencadas algumas ações como: definir tópicos e termos de pesquisa para o contexto; realizar levantamentos de trabalhos em bases de dados brasileiras; avaliar artigos coletados e excluir trabalhos dos critérios de seleção; e criar categorias analíticas para analisar o conteúdo selecionado. Nosso intento é que este estudo possa contribuir com outras iniciativas de pesquisa nessa temática. Por fim, propomos que outros estudos sejam realizados com este propósito. A revisão sistemática de literatura facilita a difícil tarefa de selecionar os dados mais importantes para o tema de pesquisa, ou seja, uma compilação de trabalhos sobre temas específicos para gerar diálogo entre pesquisadores e autores de outros trabalhos, fortalecendo, portanto, um diálogo colaborativo entre os campos da CI e da Museologia no Brasil.

Referências

BARDIN, L. **Análise de Conteúdo**. 3 reimp. 1 ed. São Paulo: Edições 70, 2016. Disponível em:https://ia802902.us.archive.org/8/items/bardin-laurence-analise-de-conteudo/bardin-laurence-analise-de-conteudo.pdf. Acesso em: 09 ago. 2022.

BRASIL. **Acervo em Rede e Projeto Tainacan**. Ministério do Turismo - Instituto Brasileiro de Museus (Ibram), 2021. Disponível em: https://www.gov.br/museus/ptbr/acesso-a-informacao/acoes-e-programas/acervo-em-rede-e-projeto-tainacan. Acesso em: 18 jul. 2022.

CATORIA, T. Ciência da Informação e museus de arte: diálogos e interações no acesso às informações do acervo do núcleo de arte contemporânea da Paraíba. In: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 11., 2010, Rio de Janeiro. **Anais eletrônicos**... Rio de Janeiro: Ancib, 2010. Disponível em: http://congresso.ibict.br/index.php/xi/enancibXI/paper/view/242/163. Acesso em: 29 ago. 2022.

IBRAM. **Cadastro Nacional de Museus**, 2010. Dados fornecidos pelo IBRAM em 14 ago. 2021 via Sistema Eletrônico do Serviço de Informação ao Cidadão, conforme Lei de Acesso à Informação (LAI), Brasil. Lei nº 12.527/2011. Acesso em: 20 ago. 2022.

IBRAM. **Museus em Números**. Brasília: Instituto Brasileiro de Museus, 2011, v. 1. Disponível em: https://www.museus.gov.br/wp-content/uploads/2011/11/Museus_em_Numeros_Volume_1.pdf. Acesso em: 22 ago. 2022.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS (IFLA). **Declaração dos Princípios Internacionais de Catalogação**. Haia, 2016. Disponível em: https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/icp/icp_2016-pt.pdf. Acesso em: 22 ago. 2022.

LANCASTER, F.W. **Indexação e resumos**: teoria e prática. Brasília: Briquet de Lemos, 2004.

LEMOS L. H.; KARPINSKI, C. CI e Museologia: análise das comunicações orais do Enancib sobre a RI. In*:* ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, n. 19, 2018, São Paulo. **Anais eletrônicos** ... São Paulo: Ancib, 2018. Disponível em: http://hdl.handle.net/20. 500.11959/brapci/102440. Acesso em: 29 ago. 2022.

PIQUET, R.; COUTO, I.; PIRES, A. A implementação da base de dados do Museu do Índio. In: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, n. 16, 2015, João Pessoa. **Anais eletrônicos** ... João Pessoa: Ancib, 2015. Disponível em: http://www.ufpb.br/evento/lti/ocs/index. php/enancib2015/enancib2015/login. Acesso em: 29 ago. 2022.

SILVA, C. A.; LARA, M. L. G. Metadados para descrição de acervos de arte no Brasil. *In*: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 20, 2019, Florianópolis. **Anais eletrônicos**... Florianópolis: Ancib, 2019. Disponível em: https://conferencias.ufsc.br/index.php/enancib/2019/paper/view/502 Acesso em: 29 ago. 2022.

_____. Esquema básico de metadados para representação descritiva de obras de arte em museus brasileiros. **Transinformação**, v. 33, 2021. Disponível em: https://www.scielo.br/j/tinf/a/DTLyDN7trqnwFcHLcLcBmQg/. Acesso em: 29 ago. 2022.

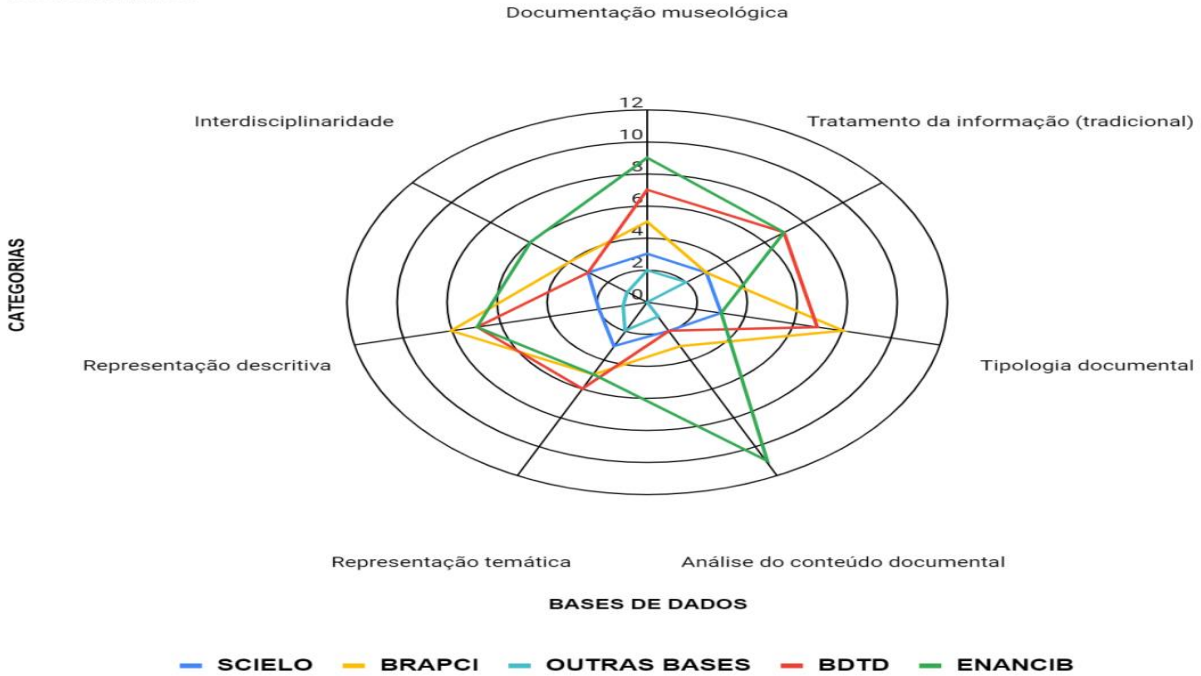
WINER, D.; ROCHA, I. E. Europeana: um projeto de digitalização e democratização do patrimônio cultural europeu. Patrimônio e Memória, São Paulo, Unesp, v. 9, n. 1, p. 113- 127, jan-jun, 2013

Apêndice A – Categorias de definições de conteúdo

Documentação Museológica- A categoria abriga informações apresentadas nos artigos referentes ao processo de musealização de objetos, bem como suas definições e etapas; também irá abordar o documento no contexto museológico.
Tratamento da Informação (tradicional)- A categoria apresenta informações na área do tratamento da informação dentro dos museus sobre os aspectos internos de classificação, descrição e representação. Portanto, foca no desenvolvimento das ações internas dentro de cada museu e suas particularidades informacionais. Trata do tratamento da informação em seu âmbito tradicional.
Tipologia Documental- A categoria abrange a descrição dos aspectos físicos e tipológicos das obras que o museu comporta e detalhes individuais adotados para a sua representação.
Análise de Conteúdo Documental- trabalha aspectos dos conteúdos desenvolvidos dentro do texto, referentes aos resultados obtidos ou observações acerca do tema. Análise sobre a metodologia desenvolvida dentro do texto.
Representação Temática- A categoria abrange informações sobre os aspectos de descrição sobre a representação temática das obras e dos termos adequados que serão utilizados. Podem ser utilizados instrumentos de organização da informação, tais como sistemas de classificação utilizados, listas de cabeçalho de assunto, tesauros, ontologias e taxonomias.
Representação Descritiva- A categoria abriga informações sobre metadados e seus padrões para a descrição de obras museais, sobre *softwares* responsáveis pela criação de repositórios e/ou bancos de dados, sobre a organização e representação da informação em seu aspecto digital, sobre os esquemas de dados e campos utilizados para a descrição das obras e Web Semântica.
Interdisciplinaridade- aborda informações sobre a relação entre a Museologia e a Ciência da Informação, as relações que essas áreas possuem e como elas trabalham em conjunto para o desenvolvimento de um trabalho mais amplo que possa comportar informações das duas áreas e que contribua para uma melhor representação e disseminação do conhecimento museológico no Brasil.

Apêndice B – Gráfico 1

GRÁFICO 1- Incidência de categorias por base de dados dos trabalhos escolhidos



Fonte: os autores

REPRESENTAÇÃO DE DADOS DE PESQUISA COM O PADRÃO DUBLIN CORE: UMA PROPOSTA DE MODELAGEM DE METADADOS PARA PROJETOS COVID-19 NO ÂMBITO DA UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

REPRESENTATION OF RESEARCH DATA WITH THE DUBLIN CORE STANDARD: A METADATA MODELING PROPOSAL FOR COVID-19 PROJECTS IN THE SCOPE OF FEDERAL UNIVERSITY OF ESPÍRITO SANTO

Beatriz Mantovani Avancini de Jesus[1], Daniela Lucas da Silva Lemos[2], Nádia Elôina Barcelos Fraga[3]

(1) Universidade Federal do Espírito Santo, Av. Fernando Ferrari, 514 - Goiabeiras, Vitória, mantovanibeatriz@hotmail.com

(2) Universidade Federal do Espírito Santo, Av. Fernando Ferrari, 514 - Goiabeiras, Vitória, daniela.l.silva@ufes.br

(3) Universidade Federal do Espírito Santo, Av. Fernando Ferrari, 514 - Goiabeiras, Vitória, nadia.e.fraga@ufes.br

Resumo

O aumento da produção científica ocasionada pela pandemia por COVID-19 trouxe a emergência de criação de sistemas de informação para a representação e recuperação da informação em ambiente digital. Esta pesquisa teve como proposta a criação de um modelo de metadados baseado no padrão *Dublin Core*, visando oferecer o acesso a dados de pesquisa, inovação e extensão para o combate à COVID-19, no âmbito da Universidade Federal do Espírito Santo (UFES). O tratamento técnico da informação incluiu princípios metodológicos das representações descritiva e temática, juntamente ao emprego do padrão de metadados *Dublin Core*, do modelo conceitual FRBR, visando a criação e modelagem de metadados no *software* Tainacan, obtendo-se como resultado dessa aplicação a criação de um repositório digital. O *Dublin Core* se mostrou um padrão de metadados flexível, visto que permite a inserção de elementos de metadados para atendimento a uma necessidade específica (como no caso do tipo de metadado "Áreas do Conhecimento"), além de permitir o emprego de modelos conceituais para a organização de requisitos funcionais, tal como o FRBR, o que promove condições favoráveis à aplicação de encontrar, identificar, selecionar e obter o objeto bibliográfico de forma mais ampla e específica.

Palavras-chave: Representação da Informação; Atividades Científicas e Tecnológicas; Padrão *Dublin Core*; Modelagem de Metadados; Repositório Digital.

Abstract

The increase in scientific production caused by the COVID-19 pandemic has brought about an emergency to create information systems for the representation and retrieval of information in a digital environment. This proposed the creation of a research model based on the Dublin Core standard, offering access to research, innovation and extension access to combat COVID-19, within the scope of the Federal University of Espírito Santo (UFES). The technical result of the principles of methodological application of the representations of descriptive and thematic treatment, executed to the work to the conceptual model of metadata *Dublin Core*, of the conceptual model FRBR as creation and modeling of data without access to this software, including the application to this software creation of digital law. Dublin Core proved to be a flexible metadata standard, as it allows the insertion of metadata elements to meet a specific need (as in the case of the "Areas of Knowledge" metadata type), in addition to allowing the models of conceptual concepts to the organization of requirements, such as the objective of finding, which promotes FRBR, the application of finding, selecting and obtaining the bibliographic object more broadly and specifically.

Keywords: Information Representation; Scientific and Technological Activities; Dublin Core Standard; Metadata Modeling; Digital Repository.

1 Introdução

O aumento da produção científica ocasionada pela pandemia COVID-19 trouxe a emergência de criação de sistemas de informação para a representação e recuperação da informação em ambiente digital. O presente estudo consistiu na construção de um repositório digital, tendo como proposta a modelagem de metadados utilizando o padrão *Dublin Core,* visando

oferecer o acesso a dados de pesquisa, inovação e extensão para o combate à COVID-19, no âmbito da Universidade Federal do Espírito Santo (UFES).

O *Dublin Core* é um padrão de metadados que descreve objetos digitais, tais como textos, vídeos, sons, imagens e sites na web facilitando a recuperação desses recursos na internet (ARAKAKI; SANTOS; ALVES, 2015; ARAKAKI; ALVES; SANTOS, 2018). O tratamento técnico da informação para fins de organização, representação e recuperação incluiu a criação e a modelagem de metadados utilizando os princípios teóricos e metodológicos da representação descritiva (GILLILAND, 2016; INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2016) e temática (FUJITA; RUBI, 2006; LANCASTER, 2004) no âmbito da Ciência da Informação. Desse modo, vislumbra-se por meio deste artigo responder a seguinte questão de pesquisa: de que maneira os princípios de organização e representação da informação podem ser aplicados na constituição de modelos de metadados consistentes para estruturar e disponibilizar acervos, repositórios e coleções digitais no ambiente em rede da Internet?

Acredita-se que os aportes teórico-metodológicos da CI, sobretudo das representações descritiva e temática, podem trazer ganhos significativos aos objetos digitais ao organizar e representar de forma consistente suas características registradas em bases de dados disponíveis na rede.

2 Objetivos

Propor um modelo de metadados baseado no padrão *Dublin Core* envolvendo coleções de dados de pesquisa, inovação e extensão, incluindo serviços tecnológicos sobre ações da UFES para o combate à COVID-19, além de produzir solução técnica para busca e recuperação da informação, fornecendo ponto de acesso único à rica informação relacionada a um tema emergente e oportuno à sociedade**.**

3 Procedimentos Metodológicos

Para a criação do repositório se adotou o *software* livre de acesso aberto Tainacan, que viabilizou a modelagem de metadados aderidos no processo de catalogação de documentos. Derivado de um projeto iniciado em 2014 em parceria com a Universidade Federal de Goiás (UFG), o Ministério da Cultura e o Instituto Brasileiro de Museus (IBRAM), a ferramenta é um *plugin* do *Software WordPress* e possui funcionalidades que auxiliam na organização e tratamento da informação, permitindo o uso de padrões de metadados e vocabulários controlados (MARTINS et al., 2017).

Também foram utilizadas ações de curadoria digital, processo que compreende a gestão dos dados de pesquisa desde a sua descoberta, planejamento e interpretação, ação que garante a preservação desses dados a longo prazo (ABREU; CONEGLIAN; VIDOTTI, 2017).

Foram selecionados quatorze projetos de extensão, integralmente acessíveis por meio do portal de projetos da Pró-Reitoria de Extensão (PROEX), oriundos da Chamada de Propostas de Projetos e Ações de Pesquisa, Inovação e Extensão para o combate à COVID-19 (Universidade Federal do Espírito santo, 2020). Após a escolha dos projetos, o procedimento seguinte consistiu na análise das características dos documentos para identificação e extração de dados relevantes (denominados requisitos funcionais) tanto para a criação do repositório digital quanto para o tratamento técnico da informação. Constaram deste procedimento de análise o uso do método de análise de assunto (FUJITA, 2003) para auxiliar no estabelecimento da correta análise conceitual entre os assuntos percebidos nos documentos.

Visando a representação e recuperação eficaz dos dados de pesquisa, os procedimentos metodológicos adotados foram subdivididos nas seguintes etapas:

a) Criação da coleção e definição de nome coerente ao contexto do repositório, denominado Projetos Covid-19 UFES.

b) Criação de vocabulários controlados, incluindo taxonomias e listas controladas.

c) Criação e modelagem dos tipos de metadados a partir do alinhamento dos requisitos funcionais com o padrão *Dublin Core*, padrão de referência global para descrição de recursos digitais.

d) Atribuição de vínculo das taxonomias criadas aos tipos de metadados modelados.

e) Catalogação dos objetos digitais (14 projetos de extensão) a partir dos tipos de metadados criados e modelados.
f) Definição e criação dos filtros de busca que consistiu na escolha dos metadados utilizados como filtros (ou facetas) para explorar o repositório em termos de navegação e buscas temáticas.
g) Validação da interface de busca (testes de usabilidade) com os próprios pesquisadores do projeto visando verificar a relevância do documento recuperado.

4 Resultados

Após a geração da coleção e definição de nome (item metodológico [a]), a criação de vocabulários controlados (item metodológico [b]) originou a taxonomia em que se estabeleceram as relações de gênero e espécie presentes no repositório. A análise das características dos documentos, item [c] da metodologia, permitiu o levantamento de metadados resultando em requisitos funcionais posteriormente alinhados com o padrão *Dublin Core* (Quadro 1).

O processo de criação dessas taxonomias consistiu na elaboração de listas controladas ou de autoridades (INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2016) inseridas no âmbito da definição dos tipos de metadados *Dublin Core* utilizados. Como um dos exemplos dessa aplicabilidade, temos a criação de listas controladas nos sites dos Centros e Departamentos da UFES, que nesta pesquisa cumpriu a função de padronizar, normalizar e facilitar a inserção de dados em um dos campos de metadados denominado Cobertura. No total foram criadas três taxonomias (Apêndice A): 1-Áreas do conhecimento; 2-Descritores; 3-Unidades de Ensino, Pesquisa e Extensão; e Órgãos Suplementares.

Taxonomias e vocabulários controlados são ferramentas importantes para o controle e padronização dos dados. Os vocabulários controlados compreendem lista de termos autorizados para representar o conteúdo de documentos indexados ou catalogados por assunto. Incluem estruturas conceituais semânticas com a função de controlar sinônimos, diferenciar homógrafos e agrupar termos com significados afins ou relacionados (LANCASTER, 2004). No contexto desta pesquisa, foram utilizados como referências à criação das taxonomias o tesauro multilíngue Descritores em Ciências da Saúde (2021) ou DeCS/MeSH, na maior parte das vezes e o Thesaurus Brasileiro de Educação (2020) ou Brased, ambos adotados no processo de indexação de assuntos.

Tratando-se da criação da taxonomia relativa aos “Descritores”, os procedimentos incluíram as etapas do processo de indexação (FUJITA; RUBI, 2006; LANCASTER, 2004). A primeira delas, análise de assunto, envolveu o exame do objeto (resumo do projeto), a identificação e a seleção dos conceitos constituídos nos resumos. A segunda etapa, tradução, envolveu a conversão dos conceitos selecionados anteriormente, em termos constituídos nos vocabulários controlados, DeCS/MeSH e Brased. Exemplificando o processo de indexação, a tradução do conceito “Materiais Educativos”, selecionado na primeira etapa é traduzido na segunda etapa para o descritor autorizado no DeCS/MeSH, “Materiais Educativos e de Divulgação”.

Desta forma, originou-se a taxonomia adaptada do DeCS/MeSH, para o descritor “Materiais Educativos e de Divulgação”, estabelecendo-se relacionamento hierárquico (noções genéricas e noções específicas).

O Quadro 1 apresenta o alinhamento entre os requisitos funcionais determinados na pesquisa e fornece um exemplo de classificação dos dados de um projeto de extensão nos elementos de metadados do padrão *Dublin Core*.

Quadro 1 – Exemplificação de alinhamento dos requisitos funcionais com os elementos *Dublin Core*

Requisitos funcionais	**Padrão *Dublin core***
1 – Autor Principal (coordenador do projeto)	**1 – Autor (Creator)**
Partelli, Adriana Nunes Moraes	Partelli, Adriana Nunes Moraes
2 – Colaboradores (Equipe do projeto)	**2 – Colaborador (Contributor)**
Santos, Aline Pestana	Santos, Aline Pestana

Santos, Isabela Lorencini Freitas, Maria Ines Dias de Coelho, Marta Pereira Marim, Thais Delabarba	Santos, Isabela Lorencini Freitas, Maria Ines Dias de Coelho, Marta Pereira Marim, Thais Delabarba
3 – Título (Título do projeto)	**3 – Título (Title)**
Produção de material educativo contendo orientações para evitar contágio e disseminação do COVID19 na comunidade Quilombola.	Produção de material educativo contendo orientações para evitar contágio e disseminação do COVID19 na comunidade Quilombola.
4 – Publicação; distribuição (Instituição de origem)	**4 – Editor (Publisher)**
São Mateus, ES: Universidade Federal do Espírito Santo. CEUNES, Departamento de Ciências da Saúde, 2020.	Universidade Federal do Espírito Santo.
5 – Localização espacial, unidade acadêmica ou administrativa, ano	**5 – Cobertura (Coverage)**
São Mateus, ES: Universidade Federal do Espírito Santo. CEUNES, Departamento de Ciências da Saúde, 2020.	Centro Universitário Norte do Espírito Santo (CEUNES), Departamento de Ciências da Saúde, São Mateus, ES, Brasil.
6 – Descrição física (Representação física)	**6 – Formato (Format)**
10 p.; URL	10 p.; URL
7 – Link de acesso	**7 – Fonte (Source)**
https://projetos.ufes.br/#/projetos/1694/informacoes	https://projetos.ufes.br/#/projetos/1694/informacoes
8 - Registro no SIEX (Identificador do documento)	**8 – Identificador de recurso (Identifier)**
nº 1694	nº 1694
9 - Início e Término do projeto	**9 – Data (Date)**
Início: 01/06/2020 ***Término:*** 01/12/2020	***Início:*** 01/06/2020 ***Término:*** 01/12/2020
10 – Natureza do recurso	**10 – Tipo (Type)**
Texto	Texto

11 – Áreas do conhecimento	**11 – Áreas do conhecimento**
Grande Área do Conhecimento: Ciências da Saúde	Grande Área do Conhecimento: Ciências da Saúde
12 - Resumo	**12 – Descrição (Description)**
OBJETIVO: Produzir e distribuir material educativo contendo orientações para evitar contágio e disseminação do COVID19 na comunidade Quilombola. METODOLOGIA: Estudo qualitativo, bibliográfico com foco na produção de cartilha, desenvolvido nas etapas: 1) Levantamento dos conteúdos científicos sobre o COVID19 pela pesquisa bibliográfica; 2) Produção de cartilha com informação sobre contágio, disseminação e como se prevenir do COVID19 na comunidade Quilombola, considerando componentes étnico-geográficos; e 3) Distribuição da cartilha nas APS. RESULTADOS ESPERADOS: Essa pesquisa irá gerar um produto - cartilha educativa que tem aplicação prática e extensionista, pois irá auxiliar os profissionais da saúde na educação em saúde com foco na redução e disseminação do vírus pela comunidade Quilombola não somente do Espírito Santo, mas também de outras comunidades Quilombolas do país pela divulgação do material pelas redes sociais.	
13 – Palavras-chaves	**13 – Descritores (Subject)**
Materiais Educativos	Materiais Educativos e de Divulgação (DeCS/MeSH)
Enfermagem	Enfermagem (DeCS/MeSH)
Promoção à Saúde	COVID-19 (DeCS/MeSH)
Saúde Coletiva	Promoção da Saúde (DeCS/MeSH)
	Educação em Saúde (DeCS/MeSH)
	Grupos étnicos (Brased)
14 – Idioma	**14-Língua (Language)**
Português	Português

Fonte: dados de pesquisa, 2021.

Após a escolha do tipo e da edição de metadados vinculados às taxonomias relacionadas no Tainacan (item metodológico [d]), partiu-se para a alimentação da base de dados do repositório (item metodológico [e]) a partir da catalogação dos metadados previamente modelados e selecionados sobre os 14 projetos de extensão, incluindo:

autor, colaborador, título do projeto, centro de ensino ao qual está vinculado, o formato do documento, fonte (URL), identificação do projeto no portal de projetos da PROEX, início e término, tipo, área do conhecimento na qual estão classificados, resumo, descritores e idioma, totalizando 14 elementos de metadados inseridos no Repositório Projeto Covid-19 UFES.

Nessa etapa também foi utilizado o modelo conceitual *Functional Requirements for Bibliographic Records* (FRBR), incluindo as entidades Obra, Expressão, Manifestação e Item (INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2016) no sentido de auxiliar na organização lógica dos metadados, a exemplo, a definição do metadado DC: *type*, elemento usado para descrever o gênero do recurso a ser catalogado.

O software Tainacan não segue um padrão de modelagem de metadados rígido para catalogação (MARTINS; LEMOS; ANDRADE, 2021), permitindo aos profissionais da informação modelar seu próprio conjunto de metadados e, sobretudo, escolher um modelo semântico, como o FRBR, que melhor os orientem na caracterização das entidades abstratas (Obra, Expressão, Manifestação e Item) por meio de atributos e relacionamentos entre elas.

Uma *Obra* é uma entidade que se encontra no nível conceitual e se define como uma criação intelectual ou artística distinta, por exemplo, a obra "Produção de material educativo contendo orientações para evitar contágio e disseminação do COVID19 na comunidade Quilombola" de Adriana Nunes Moraes Partelli" (Quadro 1). Uma *Expressão* é a realização intelectual ou artística específica assumida por uma obra, por exemplo, a obra "Produção de material educativo contendo orientações para evitar contágio e disseminação do COVID19 na comunidade Quilombola", pode ser expressa em forma de texto e registro sonoro. Uma *Manifestação* encontra-se no nível de abstração física, apesar de haver discussões na comunidade acerca de seu nível, e é a materialização de uma expressão de uma obra que pode ocorrer em livros, periódicos, filmes, animações multimídia, etc. Um exemplo de manifestação é uma edição específica de uma obra. Finalmente, uma manifestação é representada pelo *Item*, que é o objeto físico. Sendo assim, todos os projetos de extensão inseridos no repositório se expressam como textos, em formato URL. Esse metadado está ligado à categoria "Expressão" do modelo conceitual FRBR.

Após a catalogação dos projetos, a etapa seguinte (item metodológico [f]) consistiu na escolha dos metadados ou pontos de acesso estratégicos que serão utilizados como filtros abrangentes de busca e recuperação durante a exploração do repositório Projetos Covid-19 UFES no momento da navegação e buscas temáticas.

Para essa finalidade, foram aplicados os filtros: autor, idioma, cobertura, descritores, área do conhecimento. O filtro relativo à categoria *Subject* do *Dublin Core* equivale à categoria descritores utilizado como termo de busca (Apêndice B) facilitando a recuperação de informações precisas, ao ser selecionado o assunto que se pretende pesquisar.

Observa-se no apêndice B, exemplificação de pesquisa por assunto com o uso do descritor "Materiais Educativos de Divulgação". Recupera-se, por exemplo, o projeto "Produção de material educativo contendo orientações para evitar contágio e disseminação do COVID19 na comunidade Quilombola", entre outros projetos que também tratam do assunto, por autor (es) e título.

Por fim, o item [g] da metodologia buscou a validação da interface de busca com os próprios pesquisadores deste experimento, além de testes com alguns usuários do protótipo, para averiguar o quão fácil seria a navegação no sistema, bem como a aplicação dos filtros de busca, em especial os filtros com as facetas taxonômicas, que permitiram comprovar buscas, navegação e recuperação satisfatórias para os projetos inseridos no repositório.

4 Considerações Finais

Esta pesquisa abordou a proposta de um modelo de metadados baseado no padrão *Dublin Core* visando oferecer uma base de dados para viabilizar a transparência de ações de extensão para o combate à COVID-19. Frente aos resultados obtidos, a questão da pesquisa e o objetivo foram plenamente

atendidos pela aplicação de princípios teórico-metodológicos no âmbito da CI na proposição de um modelo de metadados fundamentado no padrão Dublin Core, especialmente com os aportes da representação descritiva e temática.

O *Dublin Core* se mostrou um padrão de metadados flexível, visto que permite a inserção de elementos de metadados para atendimento a uma necessidade específica (como no caso do tipo de metadado "Áreas do Conhecimento"), além de permitir o emprego de modelos conceituais para a organização de requisitos funcionais, tal como o FRBR, o que promove condições favoráveis à aplicação de encontrar, identificar, selecionar e obter o objeto bibliográfico de forma mais ampla e específica. Em se tratando de mapeamento, conforme observado no alinhamento dos requisitos funcionais (Quadro 1), os resultados se apresentaram positivos ao propósito da pesquisa, permitindo busca, navegação e recuperação de informações favoráveis no ambiente da plataforma digital Tainacan.

Referências

ABREU, Janice Pereira de; CONEGLIAN, Caio Saraiva; VIDOTTI, Silvana Aparecida Borsetti Gregório. Curadoria digital e preservação digital nos repositórios. In: Seminário em Ciência da Informação, 7, 2017, Londrina. *Anais* […]. Londrina, Paraná: Universidade Estadual de Londrina, 2017, p. 848-861.

ARAKAKI, Felipe Augusto; ALVES, Rachel Cristina Vesu; SANTOS, Plácida Leopoldina Ventura Amorim da Costa. Dublin Core: state of art (1995 to 2015). *Informação & Sociedade*: Estudos. João Pessoa, vol. 28, n.2, p. 7-20. Maio/Ago. 2018. Disponível em: https://periodicos.ufpb.br/ojs/index.php/ies/article/view/38012. Acesso em: 13 set. 2022.

_____; SANTOS, Plácida Leopoldina Ventura Amorim da Costa; ALVES, Rachel Cristina Vesu. Panorama das pesquisas sobre o padrão de metadados Dublin Core no Brasil. *Revista ACB*: Biblioteconomia em Santa Catarina. vol. 20, n. 1, p. 86-97, *Anais* [...]. Florianópolis. Jan/Abr. 2015. Disponível em: https://brapci.inf.br/index.php/res/download/89090. Acesso em: 09 set. 2022.

CAFÉ, Lígia; SALES, Rodrigo. Organização da informação: Conceitos básicos e breve fundamentação teórica. In: Jaime Robredo; Marisa Bräscher (Orgs.). *Passeios pelo Bosque da Informação*: estudos sobre Representação e Organização da Informação e do Conhecimento – EROIC. Brasília DF: IBICT, 2010, 335 p. Capítulo 6, p. 115- 129. Edição eletrônica.

Descritores em Ciências da Saúde: DeCS. São Paulo: BIREME / OPAS / OMS, 2021.

FUJITA, Mariângela S. L.; RUBI, Milena P. Um modelo de leitura documentária para a indexação de artigos científicos: princípios e elaboração e uso para a formação de indexadores. *DataGramaZero*, v. 7, n. 3, 2006.

_____. A identificação de conceitos no processo de análise de assunto para indexação. *Revista Digital de Biblioteconomia e Ciência da Informação*, Campinas, v. 1, n. 1, p. 60-90, jul./dez. 2003. Disponível em: < https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/2089 >. Acesso em: 20 nov. 2022.

GILLILAND, A. J. Setting the Stage. In: BACA, M. (Ed.). *Introduction to metadata*. 3. ed. Los Angeles: Getty Research Institute, 2016.

GRÁCIO, José Carlos A. *Metadados para a descrição de recursos da Interne*t: o padrão Dublin Core, aplicações e a questão da interoperabilidade. 2002. 127 f. Dissertação (mestrado) - Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, 2002.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS (IFLA). *Declaração dos Princípios Internacionais de Catalogação*. Haia, 2016.

LANCASTER, Frederick W. *Indexação e resumos*: teoria e prática. Brasília: Briquet de Lemos, 2004.

MARTINS, Dalton L., et al. Repositório Digital com o Software Livre Tainacan: revisão da ferramenta e exemplo de implantação na área cultural com a Revista Filme Cultura. Encontro Nacional de Pesquisa em Ciência da Informação – ENANCIB, v. 18, 2017.

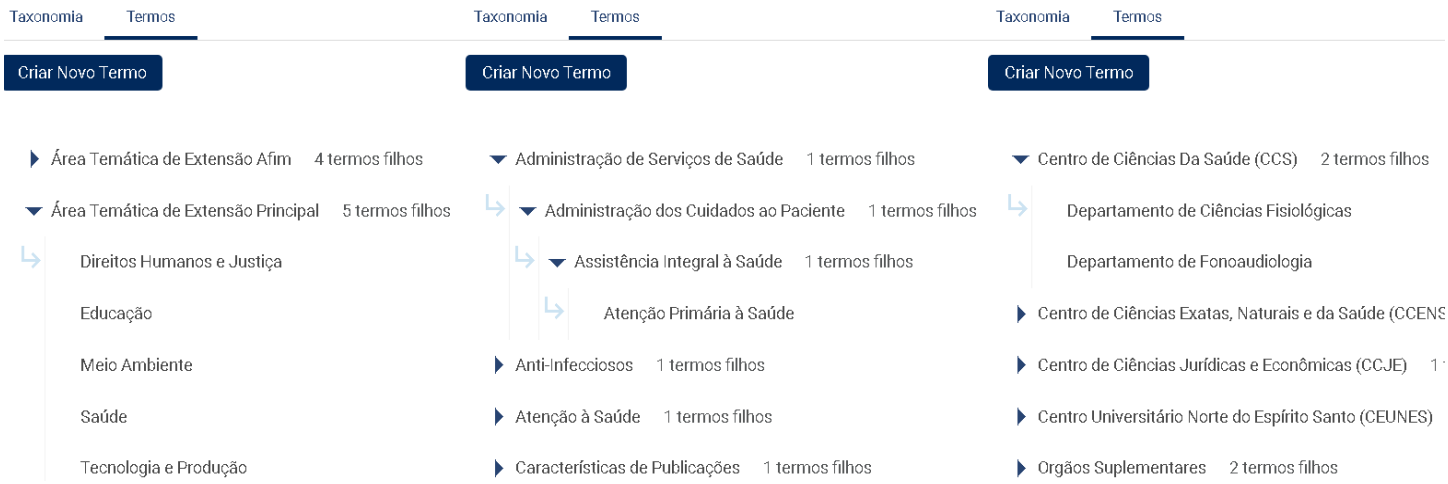
MARTINS, Dalton L.; LEMOS, Daniela Lucas da Silva; ANDRADE, Morgana Carneiro. Tainacan and Omeka: proposal for comparative analysis of software for management of digital collections from the technological effort for use and implementation. *Informação & Informação*. v. 26, n. 2, p. 569-595, 2021.

Thesaurus Brasileiro da Educação (Brased). Brasilia, DF: Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. 2020.

UNIVERSIDADE FEDERAL DO ESPIRITO SANTO. Chamada de Propostas de Projetos e Ações de Pesquisa, Inovação e Extensão para o combate à COVID-19. Vitória (ES): PRPPG; PROEX, 2020.

Apêndice A

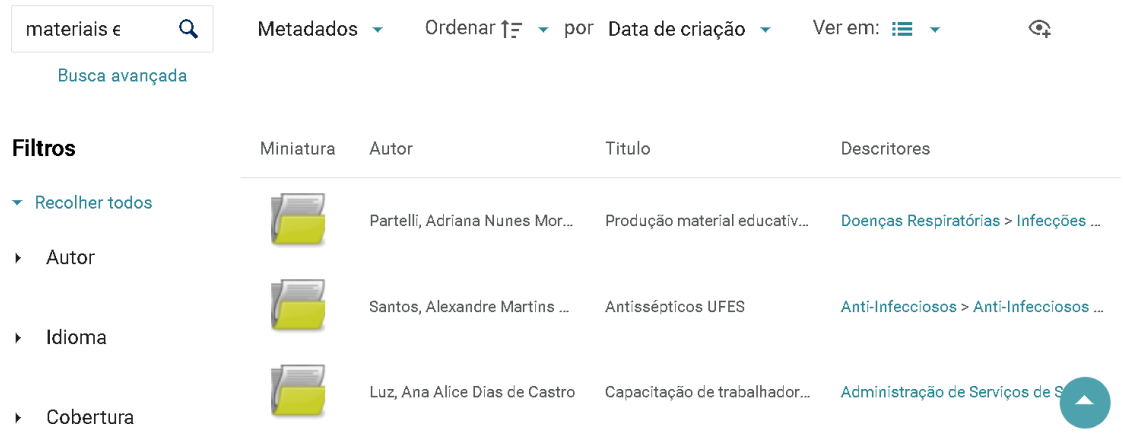
Figura 1 – Interface das taxonomias criadas no Tainacan: da esquerda para direita: 1-Áreas do conhecimento; 2-Descritores; 3-Unidades de Ensino, Pesquisa e Extensão; e Órgãos Suplementares



Fonte: dados de pesquisa, 2022.

Apêndice B

Figura 2 – Exemplificação de resultado de busca no repositório, utilizando o descritor Materiais Educativos e de Divulgação



Fonte: dados de pesquisa, 2022.

RISCOS E OPORTUNIDADES PARA OS USUÁRIOS DAS FINANÇAS DESCENTRALIZADAS

RISKS AND OPPORTUNITIES FOR DEFI USERS

Fábio Cossenzo[1], Marcello Peixoto Bax[2]
(1) Pontifícia Universidade Católica de Minas Gerais, Belo Horizonte, MG, Brasil, cossenzo@gmail.com
(2) Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil, bax@ufmg.br

Resumo
O objetivo deste artigo é identificar e analisar os principais riscos e oportunidades para os usuários das finanças descentralizadas (DeFi). O trabalho abrange o estudo da literatura sobre a arquitetura e os modelos de negócio de DeFi e a coleta de dados do Twitter para análise da percepção atual dos usuários acerca de finanças descentralizadas, com uso de técnicas de processamento de linguagem natural. Os resultados preliminares demonstram que os principais riscos são: ataques cibernéticos, custos e tempos de resposta crescentes, desvio de finalidade, dificuldade de compreensão, excesso de autonomia do usuário, irreversibilidade das transações, flutuação de preços de mercado, privacidade, restrições ao suporte e à resolução de problemas e risco de fontes externas de informação ("oráculos"). As oportunidades envolvem maior inclusão financeira, diversidade de serviços financeiros, transparência e eficiência, em comparação com as finanças tradicionais. Adicionalmente, 61% dos *tweets* analisados demonstram sentimentos positivos dos usuários com relação às finanças descentralizadas, principalmente devido à sua segurança. O principal motivo para uma percepção negativa pelos usuários parece ser a complexidade de DeFi.

Palavras-chave: Finanças Descentralizadas; DeFi; Riscos; Oportunidades; Usuários.

Abstract
This paper aims to identify and analyze the main risks and opportunities for users of decentralized finance (DeFi). The work covers the literature study on the architecture and business models of DeFi and the collection of data from Twitter to analyze the current perception of users about decentralized finance, using natural language processing techniques. Preliminary results demonstrate that the main risks are cyber-attacks, increasing costs and response times, purpose deviation, difficulty in understanding, excessive user autonomy, transactions irreversibility, market price fluctuation, privacy, restrictions on technical support and problem resolution channels and risk of external information sources ("oracles"). Opportunities involve greater financial inclusion, diversity of financial services, transparency and efficiency compared to traditional finance. Additionally, 61% of the analyzed tweets demonstrate positive feelings from users regarding decentralized finance, mainly due to its security. The main reason for a negative perception by users seems to be the complexity of DeFi.

Keywords: Decentralized Finance; DeFi; Risks; Opportunities; Users.

1. Introdução

As finanças descentralizadas, também conhecidas como DeFi – *Decentralized Finance*, prestam serviços financeiros – empréstimos, compra e venda de ativos, pagamentos etc. – em uma arquitetura financeira alternativa, aberta, implementada em *blockchains*[1] publicamente acessíveis e *permissionless*[2] (Jensen et al., 2021). Essa arquitetura se difere das finanças tradicionais ou centralizadas, em que instituições financeiras são intermediários que aproximam agentes deficitários e superavitários, em serviços centralizados de grande escala, identificando e autenticando os agentes para garantir a segurança das operações.

[1] Segundo Jensen et al. (2021), um *blockchain* é um tipo de arquitetura distribuída de banco de dados em que uma rede descentralizada de *stakeholders* mantém uma máquina de estados única. As transações representam estados de transição disseminados entre os participantes da rede em blocos de dados, cuja segurança e confiabilidade é garantida por criptografia. Um protocolo de consenso define as regras para se constituir uma transação legítima no banco de dados distribuído.

[2] *Permissionless blockchains* são ambientes abertos acessível por todos, enquanto *permissioned blockchains* são inacessíveis por partes externas não reconhecidas pelo administrador do sistema (Jensen et al. 2021).

As discussões sobre DeFi indicam o seu potencial de subjugar as finanças tradicionais e evoluir as formas atuais de contabilidade e de regulação financeira (Zetzsche, 2020). O ecossistema de DeFi (Tabela 1) está em rápida expansão (Werner et al., 2022).

Tabela 1 – Tamanho do mercado de DeFi.

Setores DeFi	**TVL[3] (USD bi)**	**Participação**
Assets	27	29%
Derivativos	23	25%
Exchanges	21	23%
Empréstimos	21	23%
Pagamentos	1	1%
Total	93	100%

Fonte: https://www.defipulse.com (em 19/09/2022).

As tecnologias envolvidas nesta transformação das finanças são resumidas no acrônimo ABCD (Zetzsche, 2020), com os seguintes benefícios: ***A****rtificial Intelligence*, eficiência e redução de custos; ***B****ig Data*, poderosa análise de dados; ***C****loud Computing*, virtualização das funções de finanças descentralizadas; e **D**LT[4], manutenção segura de registros e execução eficiente de contratos inteligentes.

Os protocolos de DeFi exigem um DLT que garanta as propriedades básicas de segurança – consistência, integridade e disponibilidade. A principal figura de DeFi é o contrato inteligente ou *smart contract*, que é um objeto executado sobre o *blockchain*, capaz de interagir com outros contratos inteligentes e garantir a consistência das transações.

Contudo, o usuário que interage com as aplicações de DeFi deve confiar inteiramente nos contratos inteligentes e na infraestrutura subjacente, os quais estão sujeitos a falhas, fraudes e ataques, de modo que os riscos e as oportunidades para os usuários não podem ser ignorados. Por outro lado, há lacunas na literatura no que se refere a identificar e analisar riscos e oportunidades do ponto de vista dos usuários de DeFi. A literatura corrente foca mais no sistema financeiro, na regulação ou nas instituições individuais.

2. Objetivos

O objetivo geral aqui é identificar e analisar os riscos e oportunidades para os usuários das finanças descentralizadas.

Os objetivos específicos traçados são:

a) Identificar riscos e oportunidades para os usuários, decorrentes da arquitetura e dos modelos de negócios de DeFi; e
b) Compreender a percepção atual dos usuários acerca dos aspectos positivos e negativos das finanças descentralizadas.

3. Procedimentos Metodológicos

Trata-se de uma pesquisa exploratória na área de DeFi ainda em fase de desenvolvimento.

O presente estudo está organizado em duas etapas. A primeira consiste na identificação de riscos e oportunidades de DeFi a partir de estudos da literatura sobre a arquitetura e os modelos de negócio, eliminando-se os aspectos que não dizem respeito ao usuário. Por exemplo, não é incluído o risco de contágio, que diz respeito à estabilidade financeira do sistema.

Na segunda etapa do estudo é realizada uma análise de sentimentos[5], visando identificar percepções positivas e negativas dos usuários acerca do tema DeFi. Essa análise inclui a construção e execução do processo disposto na Figura 1, baseado na ferramenta KNIME Analytics Platform[6] e seus componentes.

No passo 1 do processo, foram coletados 10.000 *tweets*[7] a partir de uma

[3] *Total Value Locked* (TVL) representa a soma de todos os ativos depositados em protocolos de finanças descentralizadas.

[4] DLT, ou *distributed ledger technology*, tal como o *blockchain*, é um registro imutável de transações distribuído par-a-par (Werner et al. 2022).

[5] Análise de sentimentos é uma técnica de processamento de linguagem natural, com o intuito de identificar sistematicamente percepções subjetivas dos usuários.

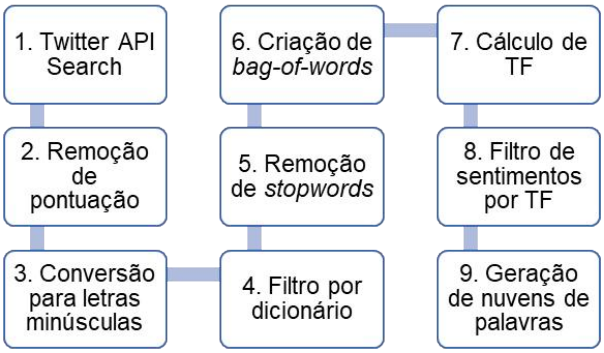
[6] https://www.knime.com/knime-analytics-platform

[7] Há uma limitação de *tweets* coletados pela Twitter API Search para o nível de acesso "elevated" (fonte: https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api).

pesquisa dos mais recentes. A consulta formulada foi a seguinte:

```
defi OR "decentralized finance" OR "decentralised finance"
```

Figura 1 – Processo de análise de sentimentos.



Fonte: Dados da Pesquisa, 2022.

Neste estudo optou-se pelo idioma inglês, porque em DeFi os termos mais utilizados ainda estão nesse idioma, mesmo no Brasil.

Os 10.000 *tweets* coletados abrangeram o período de 17/09/2022 21:42:54 até 18/09/2022 03:16:55, num intervalo de apenas 5h34min, ilustrando o grau de interesse dos usuários do Twitter pelo tema.

Nos passos 2 e 3 foram realizadas a remoção da pontuação, a separação dos textos em palavras[8] e a conversão das palavras para letras minúsculas[9].

No passo 4, foi utilizada a técnica de filtro por dicionário[10], rotulando expressões do corpus cuja subjetividade semântica pode representar opinião ou sentimento. Inicialmente, o dicionário utilizado continha 2.718 termos de conotação positiva e 4.911, de conotação negativa, todos no idioma inglês. Contudo, visando o adequado processamento dos dados coletados, os seguintes filtros adicionais foram realizados: retirados termos que haviam sido reduzidos ao seu radical (*stemmed*), pois eram duplicados no dicionário com relação aos termos derivados do mesmo radical; ignorada a definição da classe morfológica; e, finalmente, retirados termos duplicados, haja vista não haver mais a diferenciação morfológica. O dicionário passou então a contar com 1.913 termos positivos e 3.357 termos negativos.

Em seguida, nos passos 5 e 6, foi realizada remoção de *stop words*[11] e os documentos foram convertidos para *bag-of-words*[12]. Desse modo, foi possível realizar o cálculo do TF[13].

A visualização dos dados consistiu na geração de duas nuvens de palavras (passo 9) a partir dos *tweets* classificados com o sentimento "positivo" e "negativo" que foram filtrados no passo 8. O objetivo dessa visualização foi identificar os principais termos associados às percepções positivas e negativas dos usuários sobre as finanças descentralizadas.

4. Resultados

O uso adequado das finanças descentralizadas passa pelo entendimento de sua arquitetura e dos seus modelos de negócios. Na Figura 2, é mostrado um esquema de um modelo de negócio genérico de DeFi. O protocolo de DeFi na figura representa um modelo de negócio, que cobra uma taxa que é direcionada ao *DeFi's Treasury*. O serviço financeiro descentralizado é prestado pelo contrato inteligente, que aproxima investidor e usuário e garante a execução da operação de acordo com as regras de negócio estabelecidas em seu código-fonte. O investidor toma o risco do protocolo DeFi, emprestando seus recursos para a operação em troca de uma promessa de rendimentos. O usuário é quem aciona o protocolo em busca do serviço financeiro descentralizado, geralmente em tempo real, e paga taxas de juros pelo seu uso. Quem arca com os principais custos do protocolo é o usuário. Os principais modelos de negócios são empréstimos, e*xchanges*

[8] https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html

[9] https://nlp.stanford.edu/IR-book/html/htmledition/capitalizationcase-folding-1.html

[10] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

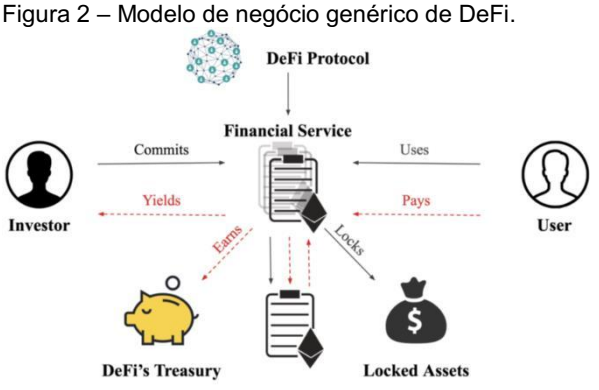
[11] *Stop words* são termos por demais frequentes e de baixa importância semântica, tais como preposições, artigos e conjunções, dentre outros.

[12] *Bag-of-words* é uma representação simplificada e eficiente para processamento de linguagem natural, que desconsidera a estrutura do texto e mantém o registro da multiplicidade das palavras.

[13] *Term frequency* (TF) é uma medida de relevância de um termo em um documento, considerando a sua frequência.

descentralizadas e agregadores de rendimentos (Xu e Xu, 2022).

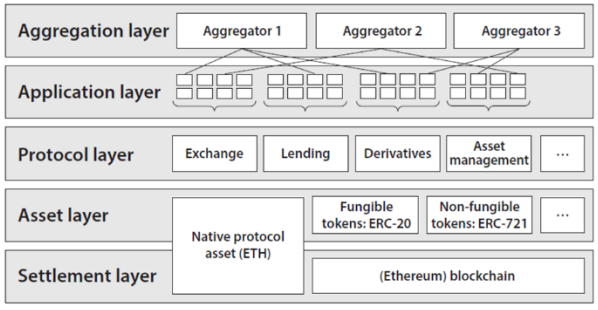
Figura 2 – Modelo de negócio genérico de DeFi.



Fonte: Xu e Xu (2022).

A arquitetura de DeFi (Figura 3) possui uma estrutura em camadas. A segurança e a confiabilidade de uma camada são diretamente dependentes das camadas inferiores (Schär, 2021). A primeira camada, *settlement layer*, consiste no *blockchain* e seus ativos nativos, tais como Bitcoin e Ethereum. Em seguida, a *asset layer* corresponde aos ativos emitidos sobre a camada anterior, também conhecidos como *tokens*. A *protocol layer* implementa os casos de uso e os contratos inteligentes. Finalmente, a *application layer* abrange os aplicativos destinados ao usuário final, enquanto a *aggregation layer* permite combinar diferentes serviços para o usuário.

Figura 3 – Arquitetura em camadas das finanças descentralizadas.



Fonte: Schär (2021).

Muitos desses contratos inteligentes necessitam dos chamados oráculos, que proveem informações de fontes externas à rede *blockchain* para que aplicações reais possam ser executadas (Schär, 2021). Por exemplo, uma transação em DeFi pode depender de uma cotação em tempo real do preço de um ativo para ser completamente efetivada por um contrato inteligente.

Decorrente da análise da arquitetura, dos modelos de negócios, dos mecanismos e dos agentes de DeFi, foram identificados na literatura consultada riscos (Tabela 2) e oportunidades (Tabela 3) para os usuários.

Tabela 2 – Riscos para usuários de DeFi.

Riscos
Ataques cibernéticos: eventuais brechas ou erros de codificação dos contratos inteligentes podem representar vulnerabilidades de segurança a serem exploradas por *hackers*, podendo resultar em desvio dos recursos dos clientes ou inutilização do protocolo.
Custos e tempos de resposta crescentes: uma maior escala de DeFi pode gerar competição excessiva pela confirmação das transações, afetando diretamente as tarifas de transação, e/ou provocar o aumento da latência na confirmação das transações que estiverem em fila.
Desvio de finalidade: aplicativos que realizam inúmeras transações em nome do cliente podem demandar uma única permissão visando a simplificação da jornada, porém esta vantagem em termos de usabilidade pode colocar em risco os fundos do cliente operados por esses aplicativos.
Dificuldade de compreensão: os contratos inteligentes possuem codificação que não é de conhecimento da população em geral, comprometendo a capacidade de o usuário tomar decisões conscientes de acordo com seus objetivos, interesses e necessidades.
Excesso de autonomia: os usuários são responsáveis pela gestão de seus ativos e pela realização de transações, pois há menor poder de atuação de autoridades centrais – reguladores, órgãos de defesa do consumidor, judiciário – sobre os contratos inteligentes e a infraestrutura descentralizada.
Irreversibilidade: se alguma transação indevida for realizada, seja por erro ou por fraude, o usuário não poderá revertê-la devido à característica do *blockchain*.
Preços de mercado: flutuações significativas podem ser observadas dos preços dos ativos, principalmente nos *tokens* secundários emitidos por protocolos DeFi.
Privacidade: visando evitar atividades ilícitas, há um potencial de os reguladores exigirem a identificação dos usuários, contudo essa medida pode afetar a privacidade dos usuários.
Restrições ao suporte e à resolução de problemas: havendo algum tipo de falha na

infraestrutura descentralizada, o usuário poderá enfrentar dificuldades e sem acesso a suporte técnico, serviço de atendimento ao consumidor, ouvidoria etc.
Risco do oráculo: as informações providas pelo oráculo podem estar equivocadas devido a falhas operacionais – no caso de um modelo centralizado –, configurar conflitos de interesses no fornecimento de informações enviesadas ou sofrer ataques de manipulação de preços para obtenção de informação favorável aos fraudadores.

Fontes: Katona (2021), Nicolas e Devos (2022), Schär (2021), Zetzsche et al. (2020).

Tabela 3 – Oportunidades para usuários de DeFi.

Oportunidades
Acessibilidade: originalmente os protocolos são *permissionless*, sem necessidade de identificar o usuário. Contudo, mesmo nos casos de exigência regulatória por identificar usuários, o controle pode ser realizado no nível dos contratos de *tokens*, sem afetar as camadas inferiores de compensação nem a característica descentralizada.
Diversificação de produtos e serviços: a arquitetura em camadas, com compartilhamento da camada de compensação, favorece a interconexão de aplicativos, possibilitando que usuários acessem uma maior diversidade de produtos e serviços, criados a partir de novas combinações, diretamente ou por intermédio de contratos inteligentes.
Eficiência: redução do risco de crédito da contraparte – portanto reduzindo as taxas de juros cobradas – e redução do tempo de realização de transações, uma vez que o processo é realizado sob a perspectiva de contratos inteligentes.
Transparência: todas as transações são publicamente observáveis e imutáveis, e os códigos dos contratos inteligentes podem ser analisados, facilitando a documentação e produção de provas.

Fontes: Katona (2021), Nicolas e Devos (2022), Schär (2021), Zetzsche et al. (2020).

Em relação à percepção atual dos usuários acerca dos aspectos positivos e negativos das finanças descentralizadas, dos 10.000 *tweets* coletados, a rotulagem de acordo com sentimentos positivos e negativos conduziu ao resultado demonstrado na Tabela 4. Note que 61% dos *tweets* representaram sentimentos positivos com relação às finanças descentralizadas, enquanto a apenas 19% foi atribuído sentido negativo.

Tabela 4 – Percepção geral dos usuários sobre DeFi

Tweets classificados	**Quant.**	**Percentual**
Sentimentos positivos	6.098	61%
Sentimentos negativos	1.876	19%
Não rotulados	2.026	20%
Total	10.000	100%

Fonte: Dados da Pesquisa, 2022.

Com relação às nuvens de palavras apresentadas na Figura 4 e na Figura 5, quanto maior é o tamanho da fonte, mais frequente é o termo. Além disso, quanto mais escura é a tonalidade da cor, mais representativo é aquele termo para o sentimento atribuído – positivo ou negativo.

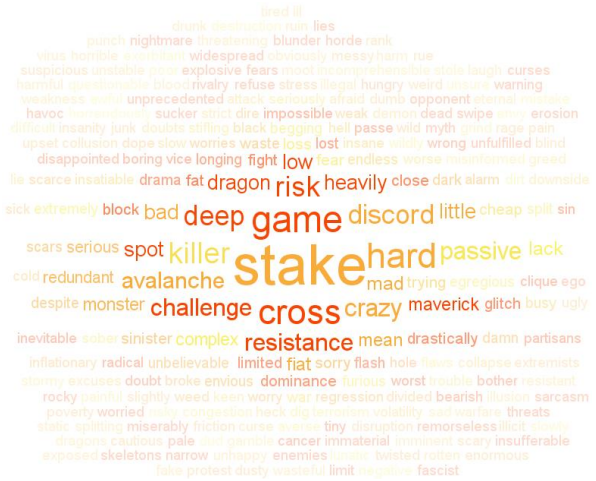
No caso da Figura 4, por exemplo, o termo “friend” é tanto o termo mais frequente quanto um dos melhores indicadores para o sentimento positivo. Já os termos “solid” e “exclusive”, por exemplo, foram melhores discriminantes do sentimento positivo do que “smart” e “secure”, os quais possuem frequências relativas similares.

Figura 4 – Nuvem de palavras associadas aos sentimentos positivos sobre DeFi

Fonte: Dados da Pesquisa, 2022.

Por outro lado, na Figura 5, por exemplo, o termo “stake” é o mais frequente, porém os termos “game”, “cross”, “deep” e “risk” foram melhores discriminantes do sentimento negativo, apesar de serem relativamente menos frequentes.

Figura 5 – Nuvem de palavras associadas aos sentimentos negativos sobre DeFi



Fonte: Dados da Pesquisa, 2022.

Com o intuito de elucidar um pouco mais o significado dos termos mais frequentes em *tweets* aos quais foram atribuídos sentimentos positivos (Tabela 5) e negativos (Tabela 6), realizou-se uma análise de co-ocorrência de termos com o propósito de melhorar a sua contextualização.

Tabela 5 – Co-ocorrência de termos relativos aos sentimentos positivos sobre DeFi.

Sentimentos positivos	Co-ocorrências
independent, smart	70
super, proud	70
top, rewarding	53
ready, secure	28
safe, secure	26

Fonte: Dados da Pesquisa, 2022.

Tabela 6 – Co-ocorrência de termos relativos aos sentimentos negativos sobre DeFi.

Sentimentos negativos	Co-ocorrências
complex, cross	9
hard, scars	7
begging, deep	6
begging, low	6
complex, hard	6

Fonte: Dados da Pesquisa, 2022.

Percebe-se, primeiramente, uma concentração maior de co-ocorrência de termos no nível da frase para os sentimentos positivos. Dentre os “top 5” para sentimentos positivos, o termo “secure” aparece duas vezes, e o termo “safe”, de sentido relacionado, aparece uma vez, denotando confiança dos usuários em DeFi. No caso de sentimentos negativos, por duas vezes aparece no “top 5” os termos “complex” e “hard”, podendo indicar que para os usuários um problema relevante é a dificuldade de uso.

5. Considerações Finais

O objetivo deste trabalho foi identificar e analisar os riscos e oportunidades para os usuários das finanças descentralizadas.

As principais contribuições deste trabalho foram a realização de uma análise de riscos e oportunidades sob o ponto de vista do usuário de DeFi e um levantamento inicial da percepção atual dos usuários do Twitter. Apesar dos diversos riscos identificados para os usuários, 61% dos tweets analisados demonstram sentimentos positivos dos usuários com relação a DeFi, principalmente devido à sua segurança.

Além dos resultados preliminares obtidos, como aprimoramentos futuros, vislumbra-se a expansão do período de análise e de outras fontes de informação, não se restringindo ao Twitter. Ainda, a análise de textos poderia incluir outros idiomas e marcadores de sentimento mais sofisticados.

Pesquisas futuras poderiam considerar legislação e regulação brasileira e internacional e analisar quais são os fatores que mais contribuem para majorar ou mitigar os riscos identificados ou explorar positivamente as oportunidades para os usuários das finanças descentralizadas.

Finalmente, um estudo mais aprofundado poderia avaliar possíveis cenários futuros de substituição parcial ou total das finanças tradicionais por DeFi, bem como seus impactos sociais, econômicos e ambientais.

Referências

JENSEN, Johannes Rude; VON WACHTER, Victor; ROSS, Omri. An Introduction to Decentralized Finance (DeFi). **Complex Systems Informatics and Modeling Quarterly**, n. 26, p. 46-54, 2021.

KATONA, Tamás. Decentralized finance: the possibilities of a blockchain “money lego” system. **Financial and Economic Review**, v. 20, n. 1, p. 74-102, 2021.

NICOLAS, Clara; DEVOS, Maxence. *Could Decentralized Finance replace Traditional Finance? A comparative analysis of the components of financial systems.* Louvain School of Management, Université catholique de Louvain, 2022. Prom.: Nguyen, Anh.

SCHÄR, Fabian. Decentralized finance: On blockchain-and smart contract-based financial markets. **FRB of St. Louis Review**, 2021.

WERNER, Sam M. et al. SoK: Decentralized Finance (DeFi). **arXiv preprint arXiv:2101.08778**, 2021.

XU, Teng Andrea; XU, Jiahua. A Short Survey on Business Models of Decentralized Finance (DeFi) Protocols. **arXiv preprint arXiv:2202.07742**, 2022.

ZETZSCHE, Dirk A.; ARNER, Douglas W.; BUCKLEY, Ross P. Decentralized Finance. **Journal of Financial Regulation**, v. 6, n. 2, p. 172-203, 2020.

TERMOS EM MOVIMENTO: DINÂMICA DA REDE TEMÁTICA NO PROJETO BIBLIOTECA COMUM

TERMS IN MOTION: DYNAMICS OF THE THEMATIC NETWORK IN THE COMMON LIBRARY PROJECT

Benjamin Luiz Franklin
Universidade Estadual de Londrina, PR, Brasil, belfra@uel.br

Resumo
Proposta: O projeto Biblioteca Comum tem coletado e selecionado Recursos Educacionais Abertos, livros, vídeos, áudios e imagens, que possam ser distribuídos em *pen drives* de baixo custo e recuperados localmente, sem a instalação de sistemas adicionais – além de um navegador padrão – para escolas sem acesso à *internet*, incrementando o acervo de suas bibliotecas. **Objetivo**: Compreender a dinâmica temática dos objetos do projeto, no curso de suas atualizações. **Metodologia**: A partir do arcabouço conceitual da Teoria Ator-Rede, iniciar análise exploratória da dinâmica de relações entre os elementos de sua rede temática diante da aplicação de algoritmos de distribuição, centralidade e agrupamento. **Resultados**: Após análise das estruturas obtidas, contrastadas à série temporal, foram localizados os eventos em que houve expansão temática no domínio. **Conclusões parciais**: Os resultados oferecem indícios de que a inclusão dos Recursos Educacionais Abertos, no modelo de negócios editorial atual, amplia as possibilidades de fomento dos acervos de Bibliotecas Escolares, além da diversidade temática acessível a professores e alunos, conforme novos atores participem desse contexto.

Palavras-chave: Dinâmica de Redes Temáticas; Recursos Educacionais Abertos; Projeto Biblioteca Comum.

Abstract
Proposal: The Common Library project has collected and selected Open Educational Resources, books, videos, audios and images, which can be distributed on low-cost pen drives and retrieved locally, without installing additional systems – beyond a standard browser – to schools without internet access, increasing the collection of their libraries. **Objective**: To understand the thematic dynamics of the project objects, in the course of their updates. Methodology: Starting from the conceptual framework of the Actor-Network Theory, start an exploratory analysis of the dynamics of relationships between the elements of its thematic network in view of the application of distribution, centrality and clustering algorithms. **Results**: After analyzing the structures obtained, contrasted with the time series, the events in which there was thematic expansion in the domain were located. Partial **Conclusions**: The results offer evidence that the inclusion of Open Educational Resources, in the current editorial business model, expands the possibilities of promoting the collections of School Libraries, in addition to the thematic diversity accessible to teachers and students, as new actors participate in this context.

Keywords: Dynamics of Thematic Networks; Open Educational Resources; Common Library Project.

1 Introdução

A Biblioteca Comum (BC) é um projeto de extensão, realizado na Universidade Estadual de Londrina (UEL), desde 2013, cujo objetivo é incrementar o acervo digital das bibliotecas escolares em busca da universalização de seu acesso, em consonância com os esforços promulgados pela Lei n.º 12.244 de 24 de maio de 2010 e a Lei Castilho, n.º 13.696, de 12 de julho de 2018, que propõem um marco legal para a promoção do livro, da leitura e da biblioteca no Brasil (BRASIL, 2010; BRASIL, 2018; FRANKLIN, 2020).

O projeto usa Recursos Educacionais Abertos (REA) (obras com fins educacionais em domínio público ou sob a licença *Creative Commons)* os quais são coletados e integrados a um sistema aberto de gestão de bibliotecas. O resultado passa por um processo de curadoria para ser distribuído em *pen drives* populares, orientado ao aprendizado contínuo dos professores e ao auxílio na preparação de suas aulas. As obras são armazenadas em um computador local, dispensando-se a necessidade de conexão à *internet*. Os usuários podem, assim, acessar o acervo por meio de sua rede interna sem fio, usando *tablets,*

celulares, ou computadores pessoais, evitando o controle das plataformas dominantes.

Tal projeto é constituído de quatro tarefas contínuas, quais sejam: coleta, em que se busca objetos digitais sob licenças abertas, armazenado-os em diretório específico; curadoria, em que parte do material coletado é indexado e integrado a um sistema de recuperação de informação – os buscadores; catalogação, em que profissionais da informação catalogam os recursos de todos os buscadores, os quais geram um único banco de dados integrado a um sistema de gestão de bibliotecas – o biblivre; e a distribuição, em que se propõe o uso prático dessa tecnologia como ferramenta pedagógica em disciplinas no departamento de Ciência da Informação da UEL.

A base de dados atual do projeto Biblioteca Comum contém mais de 190.000 objetos digitais abertos em 4,42 TB de dados discriminados, aproximadamente, em: 55.000 textos (125 GB), 33.600 vídeos (4,7 TB), 3.500 áudios (24 GB) e 98.600 imagens (102 GB).

Na fase de curadoria são compilados os buscadores, repositórios portáteis de arquivos selecionados, conjugados mediante critério de relevância e ocupação de espaço no dispositivo, com o intuito de se ofertar o maior número de arquivos importantes, em um menor espaço possível. Destarte, tais sistemas foram pensados para serem os mais fáceis de se difundir, copiar e utilizar, sem que seja necessário a instalação de *softwares*, além do navegador de *web*, e por possuir baixíssimo custo de manutenção. Ademais, destaca-se, no presente trabalho, o buscador de textos (BC-TEXTO 2.2.7), o qual possui acervo de 4.173 livros clássicos e científicos, em ciências sociais e humanidades.

Nessa seara, nos últimos dez anos, mais de 220 instalações foram concluídas em escolas públicas no entorno da UEL, cujos temas dos textos coletados foram, presumivelmente, ampliados, conforme mais livros tenham sido adicionados ao diretório do projeto, no decorrer de sucessivas versões[1]. Entender a dinâmica da expansão temática da coleção se faz de suma importância à orientação do projeto, ao direcionamento de seu público-alvo e adequação de futuras versões.

2 Objetivos

Este trabalho inicia uma análise exploratória da rede temática do buscador de textos do projeto Biblioteca Comum, verificando-se sua possível expansão de elementos no decorrer de suas versões e adição de novos objetos à coleção.

3 Procedimentos Metodológicos

A Teoria Ator-Rede (TAR), a qual surgiu nos anos 1980, pode ser entendida como uma abordagem de pesquisa que, em síntese, ocupa-se da compreensão das relações entre atores em contextos diversos, focando as associações, os agrupamentos e as repetições de objetos agregados em seus aspectos descritivos (LATOUR, 2005).

Dada sua ênfase nas relações materiais de comunicação entre atores, entidades humanas ou não humanas, que promovem uma transformação na rede, ganhou notoriedade em diversos tipos de estudos tais como a antropologia, sociologia, ciência e tecnologia, e, mais recentemente, com a transformação digital acelerada das últimas décadas, despontou como um poderoso ferramental teórico nos estudos dos fenômenos contemporâneos (WILLIAMS, 2022).

A inclinação etnometodológica da TAR, com foco a assimilar relações locais sociomateriais de comunicação entre atores, a qual visa dispensar uma perspectiva totalizante dos contextos pesquisados, tem sua utilização tendente a um repositório de conceitos integrados em detrimento de uma teoria explicativa totalizante (FROHMANN, 1995; JENSEN, 2019). Apesar dessas observações, muitas vezes, críticas, a TAR vem se consolidando como um instrumento de compreensão das relações sociotécnicas nos ambientes digitais, principalmente por já considerar, desde sua tenra concepção, os arranjos tecnológicos inseparáveis dos

[1] Mais informações sobre o projeto Biblioteca Comum podem ser encontradas no endereço: http://bibliotecacomum.com.br

ambientes sociais e seus dispositivos de ação e normatização.

Essas características fundamentais, pensadas décadas antes da radicalização da transformação digital no mundo pós-pandêmico, por assim dizer, preparou, antecipadamente, a TAR a instrumentalizar conceitualmente problemas relacionais entre atores: uma espécie de síntese teórica *avant la lettre* dos diversos problemas de análise da ciência de dados atuais, cujas agregações, relações semânticas, correlações e mecanismos de predição, podem resumir parte significativa dos problemas nos ambientes informacionais hodiernos (WEINBERGER, 2011).

Essa abordagem relacional entre atores, assim como a consideração de que esses são actantes ativos, transformadores de atributos de outros objetos, predispôs, desta forma, o uso da TAR em contextos documentais, em que esses são considerados expressões comunicacionais sociomateriais (PONTILLE, 2019).

Dada a convergência documental para a materialidade digital hegemônica da última década, a unificação representacional da relação entre atores contidos nos objetos digitais concluiu o que Morozov (2018) denomina de hiper inclusão, ou seja, a produção de um ambiente global unificado, baseado na máquina universal computacional que, ao operar seus objetos a partir de uma materialidade hegemônica, unifica, também, sua relação contextual, colocando-se, assim, todos os atores em um mesmo domínio informacional, transformando a dinâmica de mudança de seus atributos em operações contábeis e ativos financeiros a serem geridos, uma grande fábrica contemporânea que inclui a todos (FRANKLIN, 2020).

Essa unificação da disponibilização de documentos**,** por meio de objetos digitais, exponencializou o tratamento analítico dos conteúdos, tornando a tradição de estudos, já em curso desde o início do século XX, completamente compatível com o arsenal de técnicas computacionais contemporâneas, como a teoria de grafos, análises estatísticas e modelos preditivos (KRIPPENDORFF, 2004).

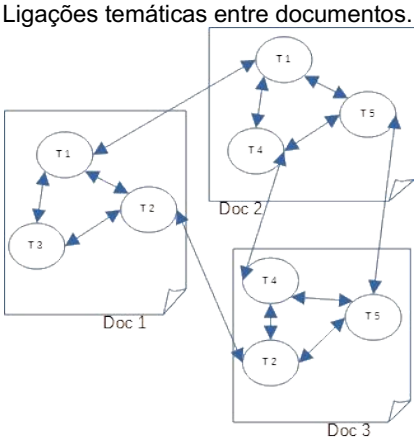
A tradição analítica dos estudos nas Ciências Documentais, de domínio, métricas, temáticas, documentos, dentre outras, ganharam um novo arcabouço de técnicas de análise, seja ao estilo indutivo, analítico, *bottom up*, técnicas utilizadas usualmente nos sistemas computacionais, conquanto diferentes dos modelos hierarquizados e dedutivos, típicos dos tradicionais esquemas de classificação (CAMPOS, 2018).

Nesse contexto, abordar-se-á o problema da evolução temática no buscador de textos da Biblioteca Comum, como uma evolução de uma rede sociomaterial, passível de ser analisada enquanto um grafo dinâmico representante de uma relação entre atores. Os temas serão tratados como representantes de uma rede material de relações terminológicas, que pode ser distribuída, agrupada e analisada pelos algoritmos comuns à análise de grafos (PINHEIRO, 2019).

Para tanto, considerou-se os termos de indexação extraídos dos dados catalográficos de cada livro, como nodos de uma rede, cuja presença material em cada objeto digital adicionou uma relação, ou seja, arestas entre os nodos temáticos, unificando-se os 4.173 livros a partir de temas comuns, os quais formaram um grafo de 5.577 vértices de termos distintos e 12.347 arestas, associados à linha temporal gerada pelas datas de alteração dos objetos digitais, resultando um grafo dinâmico (LANUM, 2016).

Dessa forma, a presença material de Temas (T), como nodos discretizados da rede, no mesmo livro (Doc), indica uma relação de latência semântica entre esses termos, em que o sentido é apenas suposto e presumido, apesar de não estar formalizado e explícito, o qual produz uma relação bidirecional entre os nodos do grafo (HOFMANN, 2013). Ademais, a presença de igual termo em documentos diferentes produz uma relação bidirecional entre esses nodos, os quais ligam os diferentes documentos a um tema comum, conforme a figura 1.

Figura 1: Ligações temáticas entre documentos.



Fonte: Dados da Pesquisa, 2022.

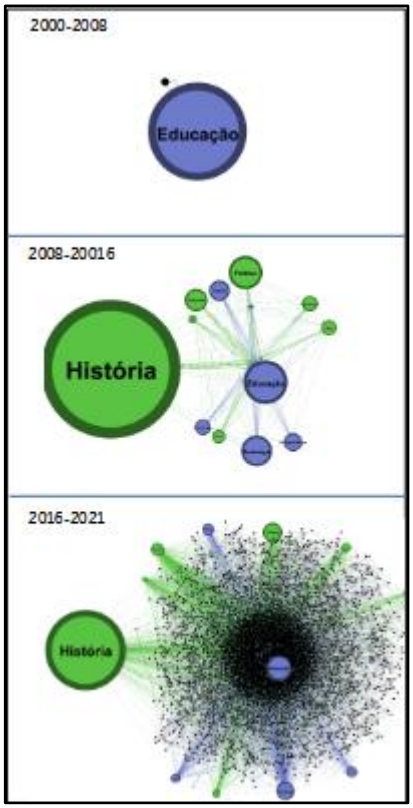
Essa forma de indexação pós-coordenada, conforme Lancaster (2004), aborda os documentos de maneira multidimensional, considerando a ligação entre seus diversos temas como uma rede não hierarquizada de relacionamentos, que os contextualizariam.

4 Resultados

Após a implementação do grafo por meio do pacote *Gephi* (BASTIAN *et al.*, 2009), foram aplicados os algoritmos de centralidade de intermediação e grau, além da detecção de grupos, associados à linha do tempo obtida pela extração da data de modificação dos objetos digitais da coleção, formando-se um grafo dinâmico. Nomearam-se os grupos conforme o nodo respectivo mais relevante na centralidade de intermediação, obtendo: *educação*, *história* e *literatura*, cuja dinâmica aponta três momentos principais na história do projeto, a saber: fase 1, entre 1999 até 2008; fase 2, entre 2008 até 2016; e fase 3, entre 2016 e 2021, conforme a figura 2.

Nesse sentido, faz-se importante destacar não haver, necessariamente, adesão disciplinar, ou subordinação entre o nome dos grupos e outros termos participantes, tais como: *educação e tecnologia*, ou *história e política*, observando-se apenas uma aproximação temática dada sua latência semântica. Essa possível relação de subordinação conceitual não está no escopo do presente trabalho, não obstante, poderá ser objeto de investigações futuras.

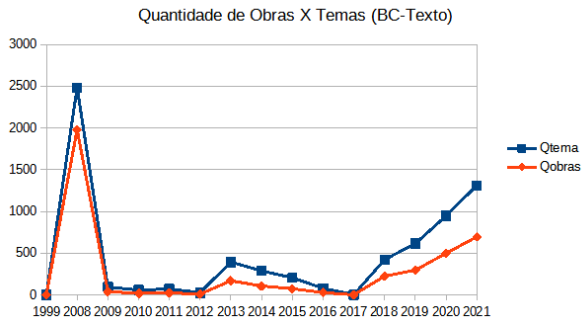
Figura 2: Três momentos na dinâmica.



Fonte: Dados da Pesquisa, 2022.

Por conseguinte, a agregação das classes formadas contrastadas à linha do tempo mostra picos de concentração temática, em 2008, 2013 e um reforço constante a partir de 2017, conforme a figura 3.

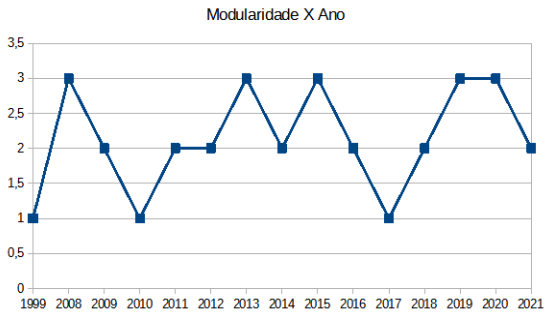
Figura 3: Documentos e Temas.



Fonte: Dados da Pesquisa, 2022.

Se o evento de 2008 aponta para uma estruturação do campo temático, conforme a figura 4, somente de 2017 em diante, nota-se o aumento do número de obras sem, contudo, alterar a quantidade de *clusters* já estabelecida previamente.

Figura 4: Número de classes formadas.



Fonte: Dados da Pesquisa, 2022.

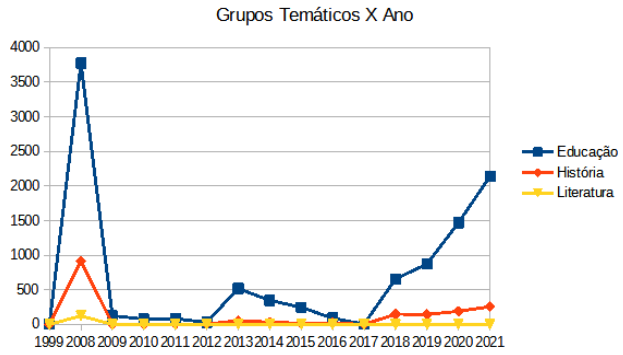
Destarte, o crescimento gradual da expansão temática, a partir de 2017, parece indicar uma diferença nas condições de inclusão de Recursos Educacionais Abertos nos circuitos editoriais, dada a diferença do primeiro evento em 2008, mais abrupto e temporário, e 2013, ainda ensaístico. Se, no evento de 2008, observa-se a explosão seminal de campos temáticos, em 2017, verifica-se plausível uma consolidação da adesão de novos atores no cenário da educação aberta, como sugere Carvalho (2021).

Uma possível explicação seria a adesão de licenças abertas ao modelo de negócios de editoras comerciais. Ao usar o *Creative Commons*, pequenas editoras fragmentariam e ampliariam o uso de licenças abertas, enquanto na década anterior, apenas grandes iniciativas governamentais centralizadas, como o Portal Domínio Público, popularam pioneiramente a *internet* com material aberto (PIRES *et al.*, 2016). Essa é, no entanto, uma discussão extensa que não se enquadra no escopo deste trabalho.

A retomada da expansão temática a partir de 2017, apesar de constante, não configura, aparentemente, mais grupos que os formados em 2008, conforme a figura 4.

Essa qualidade pode ser uma característica do próprio nicho temático do projeto Biblioteca Comum, ou indicativo de amadurecimento do próprio domínio, no sentido de se estabelecer uma consolidação nas principais áreas temáticas, conforme a figura 5.

Figura 5: Documentos em áreas temáticas.



Fonte: Dados da Pesquisa, 2022.

5 Conclusões Parciais

Esse trabalho buscou iniciar uma análise exploratória na coleção do projeto Biblioteca Comum, com o intuito de se identificar os principais eventos de sua presumida expansão temática, ao longo de sucessivas versões na última década. A compreensão da dinâmica dessa rede temática habilita melhores decisões quanto ao seu desenvolvimento, além de adequar de modo mais preciso a coleção a seus usuários. Por conseguinte, os resultados preliminares obtidos a partir da análise do grafo de sua rede temática, indicam que os principais grupos foram criados já em 2008, embora tenham alcançado, somente em 2017, um crescimento constante.

Nesse sentido, tal característica pode indicar a entrada de novos atores no cenário editorial, antes dominado por iniciativas governamentais, que usariam licenças abertas para incrementar sua rede temática, ampliando o público compatível com seus modelos de negócios. O resultado dessa expansão de títulos, preservando as categorias mais genéricas de agrupamento, indicaria uma possível multiplicação de especializações temáticas das obras, no decorrer das últimas versões do buscador de textos. Os resultados desse trabalho, no entanto, apenas contribuem para essa conjectura, indicando um espaço aberto para futuras pesquisas. A diversidade temática e o alcance desses recursos, de outra forma, são evidenciáveis ao apontarem uma alternativa valiosa à tarefa de enriquecimento das bibliotecas escolares.

Referências

BASTIAN, Mathieu; HEYMANN, Sebastien; JACOMY, Mathieu. Gephi: An Open Source

Software for Exploring and Manipulating Networks. **Proceedings of the International AAAI Conference on Web and Social Media**, v. 3, n. 1, p. 361–362, 2009.

BRASIL. Presidência da República. Casa Civil. Subchefia para assuntos Jurídicos.**Lei n.º 12.244, de 24 de maio de 2010**. Dispõe sobre a universalização das bibliotecas nas instituições de ensino do País. Disponível em:http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2010/lei/l12244.htm. Acesso em: 13 de nov. 2022.

BRASIL. Presidência da República. Casa Civil. Subchefia para assuntos Jurídicos. **Lei n.º 13.696, de 12 de julho de 2018**. Institui a Política Nacional de Leitura e Escrita. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13696.htm. Acesso em: 13 de nov. 2022.

CAMPOS, Maria Luiza de Almeida. Modelização de domínios de conhecimento: uma investigação de princípios fundamentais. **Ciência da Informação**, v. 47, n. 1, 2018.

CARVALHO, Tassiana Cunha. **A adoção do modelo de Recursos Educacionais Abertos no Programa Nacional do Livro e do Material Didático**. 2021. Disponível em: https://repositorio.unb.br/handle/10482/43441 Acesso em: 8 nov. 2022.

FRANKLIN, Benjamin Luiz. A recusa à escassez: a universalização da biblioteca escolar e a tensão entre a materialidade digital e a Lei de Direitos Autorais. **Informação & Sociedade: Estudos**, v. 30, n. 3, p. 1–23, 2020.

FROHMANN, Bernd. Taking Information Policy beyond Information Science: Applying the Actor Network Theory. *In*: **Connectedness: Information, Systems, People, Organizations, the 23rd Annual Conference of the Canadian Association for Information Science**, 1995.

HOFMANN, Thomas. **Probabilistic Latent Semantic Analysis**. 2013.

JENSEN, Casper Bruun. Is actant-rhizome ontology a more appropriate term for ANT? In: **The Routledge Companion to Actor-Network Theory**. Routledge & CRC Press, 2019.

KRIPPENDORFF, Klaus. **Content analysis: an introduction to its methodology**. 2nd ed. Thousand Oaks, Calif: Sage, 2004.

LANCASTER, F. W. **Indexação e resumos: teoria e prática**. 2. ed. rev. atual. Brasília, DF: Briquet de Lemos/Livros, 2004.

LANUM, Corey L. Dynamic graphs: how to show data over time. In: **Visualizing Graph Data**. Manning Publications, 2016.

LATOUR, Bruno. **Reassembling the social: an introduction to actor-network-theory**. Oxford; New York: Oxford University Press, 2005.

MOROZOV, Evgeny. **Big Tech**. Ubu Editora, 2018.

PINHEIRO, Cristiano Guedes. **Uma análise sobre a utilização de dados coletados nas redes sociais online para a pesquisa acadêmica e os procedimentos contidos nos manuais de metodologia científica**, 2019.

PIRES, Erik André de Nazaré; GIRARD, Carla Daniella Teixeira; GIRARD, Cristiane Marina Teixeira. O Portal Domínio Público como auxílio tecnológico no escopo da pesquisa no século XXI. **Múltiplos Olhares em Ciência da Informação**, v. 6, n. 1, 2016.

PONTILLE, Jérôme D. What did we forget about ANT's roots in anthropology of writing? *In*: **The Routledge Companion to Actor-Network Theory**. Routledge & CRC Press, 2019.

WEINBERGER, David. **Too big to know: rethinking knowledge now that the facts aren't the facts, experts are everywhere, and the smartest person in the room is the room**. New York: Basic Books, 2011.

WILLIAMS, Idongesit. Contemporary Application of Ant: An Introduction. *In*: WILLIAMS, Idongesit (Org.). **Contemporary Applications of Actor Network Theory**. Singapore: Springer, 2020.

TRATAMENTO DA INFORMAÇÃO EM ACERVOS CULTURAIS: AVALIAÇÃO DO USO DE VOCABULÁRIOS CONTROLADOS EM COLEÇÕES MUSEOLÓGICAS SOB GESTÃO DO INSTITUTO BRASILEIRO DE MUSEUS

TREATMENT OF INFORMATION IN CULTURAL COLLECTION: EVALUATION OF THE USE OF CONTROLLED VOCABULARIES IN MUSEUM COLLECTIONS MANAGED BY THE BRAZILIAN INSTITUTE OF MUSEUMS

Abeil Coelho Júnior[1], Daniela Lucas da Silva Lemos[2]
(1) Universidade Federal do Espírito Santo, Vitória-ES, abeilc@hotmail.com
(2) Universidade Federal do Espírito Santo, Vitória-ES, daniela.l.silva@ufes.br

Resumo
Nos últimos anos, tem-se observado considerável adesão das instituições de patrimônio cultural ao processo de digitalização e disponibilização de seus dados de acervos na internet, proporcionando maior acesso e democratização de conhecimento científico e cultural à sociedade. Porém, apenas disponibilização de dados *online* não é o bastante atualmente, considerando que a qualidade desses dados deve ser mensurada. Assim, este trabalho busca avaliar a adequação das coleções museológicas disponibilizadas *online* pelo Instituto Brasileiro de Museus acerca do uso de vocabulário controlado, à luz das orientações do guia de catalogação de objetos culturais de referência, o *Cataloging of Cultural Objects*. Para tal, foi realizado o alinhamento dos metadados utilizados pelo Instituto Brasileiro de Museus e o guia de catalogação, desenvolvido *script* em Python para processamento dos dados das coleções, além de cálculo de adequação a recomendação de uso de vocabulário controlado. Os resultados demonstram que apenas 5 (cinco) dos 9 (nove) elementos recomendados fazem uso de vocabulário controlado; apresenta ainda que, apesar desses museus e coleções estarem sob gestão do Instituto Brasileiro de Museus, as práticas de catalogação não estão alinhadas entre as instituições. Conclui-se que práticas de catalogação provenientes de instrumentos de referência sejam incorporadas na modelagem de metadados das bases de dados dos museus sob gestão do Instituto Brasileiro de Museus, visando qualificar seus atuais padrões de documentação.

Palavras-chave: Museus. Organização da Informação. Representação da Informação. Qualidade de dados. Vocabulário controlado.

Abstract
In recent years, there has been considerable adhesion of cultural heritage institutions to the process of digitizing and making their collection data available on the internet, providing greater access and democratization of scientific and cultural knowledge to society. However, just making data available online is not enough nowadays, considering that the quality of this data must be measured. Thus, this work seeks to evaluate the adequacy of museum collections made available online by the Brazilian Institute of Museums regarding the use of controlled vocabulary, in the light of the guidelines of the cataloging guide for cultural objects of reference, the Cataloging of Cultural Objects. To this end, the alignment of the metadata used by the Brazilian Institute of Museums and the cataloging guide was carried out, a script in Python was developed for processing the data from the collections, in addition to calculating the suitability for the recommendation of using a controlled vocabulary. The results show that only 5 of the 9 elements recommended use controlled vocabulary; also shows that, despite these museums and collections being managed by the Brazilian Institute of Museums, cataloging practices are not aligned between the institutions. It is concluded that cataloging practices from reference instruments are incorporated into the metadata modeling of the databases of museums managed by the Brazilian Institute of Museums, in order to qualify their current documentation standards.

Keywords: Museums. Information Organization. Information Representation. Data quality. Controlled vocabulary.

1. Introdução

Nos últimos anos, tem-se observado considerável adesão das instituições de patrimônio cultural ao processo de digitalização e disponibilização de seus dados de acervos na internet, proporcionando maior acessibilidade e democratização de conhecimento científico e

cultural à sociedade. Entretanto investir no processo de digitalização e disponibilização de objetos culturais não é suficiente (MARTINS et al., 2022), visto que questões acerca da qualidade de dados nesses processos frequentemente não são levantadas, considerando os diversos tipos de bancos de dados e sistemas de informação ora envolvidos em processos de organização, modelagem e representação.

Diante deste cenário, dados se tornam cada vez mais recursos importantes e valiosos para o século XXI. Dessa forma, considerações acerca da importância da qualidade para publicação de conjunto de dados na internet surgem em diversos contextos (BIZER; HEATH; BERNERS-LEE, 2009; WILKINSON et al., 2016; MACEDO; LEMOS, 2021; MARTINS et al., 2022).

Para o contexto da presente pesquisa, custodiadores e proprietários de dados, como, por exemplo, galerias, bibliotecas, arquivos e museus – GLAMs, acrônimo em inglês para tais termos, são os principais responsáveis pela qualidade de seus dados, com uma boa catalogação descritiva e de assunto (IFLA, 2016). Com o uso de padrões de documentação que orientam a estrutura de dados, valores de dados e conteúdo de dados (GILLILAND, 2016), as instituições contam com um conjunto de ferramentas que pode levá-las a uma boa prática de catalogação, documentação consistente, e, por consequência maior acesso aos documentos pelo usuário final. No entanto, aqueles que fornecem os dados e aqueles que usam os dados também têm responsabilidades. Os coletores de dados e catalogadores têm o dever de rotular os dados corretamente e documentar metodologias de captação; os custodiadores têm o papel de fazer a manutenção e o controle de qualidade dos seus registros; e os usuários em reportar eventuais erros encontrados (CHAPMAN, 2005).

Com objetivo de viabilizar a interoperabilidade, acesso e reúso de seus dados, instituições de patrimônio cultural (HARPRING, 2022) geralmente aderem a padrões de catalogação que produzem descrições de coleções de objetos culturais digitalizados ou nato digitais, os quais necessitam padronizar aspectos únicos de coleções culturais tanto fisicamente quanto digitalmente, bem como fornecer dados administrativos para descrever a digitalização, os direitos autorais e as disposições de uso dos objetos.

Nesse sentido, como forma de garantir a qualidade de dados de uma coleção cultural, faz-se necessário a adoção de padrões de tratamento documental atrelados ao uso de vocabulários controlados.

Vocabulário controlado é definido por Lancaster (2004) como uma lista de termos autorizados, em que o catalogador ou indexador somente pode atribuir a um documento termos presentes na lista adotada pela unidade de informação envolvida.

Os vocabulários controlados podem auxiliar: i) nos processos de análise e descrição de documentos, permitindo a criação padronizada de metadados ao nomear, de forma consistente, os pontos de acesso aos documentos e a informação neles contida; e ii) no processo de busca em um sistema de recuperação de Informação através da padronização e expansão do vocabulário controlado das consultas. Exemplos de vocabulários controlados incluem: esquemas de classificação, listas de cabeçalhos de assuntos, tesauros, taxonomias e ontologias (ABBAS, 2010).

Como padrões de tratamento documental, Abbas (2010) destaca dois desses padrões que vislumbram a produção de bases de dados no âmbito do patrimônio cultural, especialmente para museus, a saber: i) o padrão semântico *Categories for the Description of Works of Art* (CDWA) e sua extensão *Cataloging of Cultural Objects* (CCO); e ii) o padrão de metadados *Visual Resources Association Core Categories* (VRA Core).

O CCO foi publicado em 2006, como resultado do consenso de profissionais das comunidades de museus, bibliotecas, galerias e arquivos que pesquisam a prática comum entre essas disciplinas (BACA et al., 2006, HARPRING, 2022) e fornece diretrizes para selecionar, ordenar e formatar dados usados para preencher registros de catálogo com base em categorias principais em CDWA e VRA Core. Porém, embora tenha sido inspirado no desenvolvimento dos elementos VRA Core e do *Getty Vocabularies*, o CCO apresenta conceitos mais genéricos que podem ser utilizados com

outros conjuntos de metadados, como, por exemplo, o *Machine-Readable Cataloging* (MARC21), *Metadata Object Description Schema* (MODS), Dublin Core, entre outros.

No caso do Instituto Brasileiro de Museus (Ibram), suas bases de dados foram modeladas a partir do padrão de dados adotado internamente pela instituição, qual seja o modelo do Inventário Nacional de Bens Culturais Musealizados – INBCM (BRASIL, 2021).

2. Objetivos

O objetivo geral desta pesquisa é fazer avaliação da qualidade de dados na perspectiva do uso de vocabulário controlado nos elementos de metadados que descrevem as coleções dos museus sob gestão do Ibram.

3. Procedimentos Metodológicos

O INBCM possui elementos discricionais para coleções museológicas, bibliográficas e acervos arquivísticos. Porém, para os objetivos deste trabalho foram considerados apenas os elementos de caráter museológico. Desta forma, o primeiro passo foi realizar o alinhamento (mapeamento) entre os elementos descritivos de carácter museológico do INBCM com os elementos recomendados pelo guia de catalogação CCO, conforme Apêndice A. O alinhamento se deu a partir de um procedimento manual e intelectual baseado na aquisição de conhecimento sobre os dois instrumentos de pesquisa.

O segundo passo, a partir do alinhamento, foi captar os dados do Ibram a partir de *script* (COELHO, 2022) por meio da utilização da linguagem de programação Python, e com o uso das bibliotecas *Pandas*[1], *BeautifulSoup*[2] e *Requests*[3], para realizar a exportação em massa de todos os dados dos acervos dos museus no formato "CSV: inbcm-ibrammapper" do Tainacan. Sabendo-se que a ferramenta de repositório usada pelos museus sob gestão Ibram adota vocabulários controlados do tipo taxonomia na modelagem de seus metadados, fez-se a verificação do seu emprego em cada elemento de metadado, alinhado com o CCO, com exigência de uso de vocabulário controlado. A avaliação foi feita a partir dos dados disponibilizados pela API do próprio Tainacan, disponível no painel de exportação com nome "API do Tainacan em formato JSON". Essa API disponibiliza dados para além dos elementos de metadados do INBCM, dentre eles a indicação da configuração do elemento de metadado caso este seja do tipo taxonomia em uma determinada coleção.

Por fim, foram captados os dados de todas as 22 coleções de caráter museológicos disponíveis online pelo Ibram. E avaliados se os elementos de metadados nestas coleções possuíam a indicação de taxonomia. Desta forma, para cada coleção e elementos de metadados que possuíssem a indicação de taxonomia e que tivesse valor preenchido foi atribuído: (i) o valor 1 (um), por estar de acordo com a regra de uso de vocabulário controlado, conforme regra do CCO; ou (ii) o valor 0 (zero), por não atender a recomendação do CCO, seja pelo elemento de metadado não ter a configuração de taxonomia ou não apresentar valor preenchido. Assim, para todas as coleções e elementos de metadados foram calculados a taxa de adequação a regra de uso de vocabulário controlado do CCO com a seguinte formula:

$$\text{índice}_b = (\sum \text{Valor1} / (\sum \text{Valor1} + \sum \text{Valor0})) * 100$$

Onde: "b" é a base com dados de uma coleção em particular; **índice** é o percentual de adequação obtido a nível de dimensão, elemento de metadado e regra de catalogação para um determinado museu e coleção; **Valor1** é a indicação de ocorrência do registro de dado que **se adequou** a regra; **Valor0** é a indicação de ocorrência do registro de dado que **não atendeu** a regra.

4. Resultados

A luz do alinhamento entre elementos de metadados do INBCM e CCO (Apêndice A) pode-se observar pela coluna "Vocabulário Controlado CCO" que para 9 (nove) dos 15 elementos são recomendados o uso de vocabulário controlado pelo CCO.

[1] https://pandas.pydata.org

[2] https://beautiful-soup-4.readthedocs.io/

[3] https://requests.readthedocs.io/en/master/

A taxa de adequação dos elementos de metadados das bases de dados do Ibram que tiveram a indicação de uso de taxonomia pode ser observada no Apêndice B. Observa-se nas linhas, os elementos de metadados alinhados com regras de catalogação de uso de vocabulário controlado e nas colunas as coleções museológicas disponíveis online pelo Ibram. Cada célula apresenta a taxa de adequação das coleções à regra de uso de vocabulário controlado, com valores de 0 (zero), onde foi observado a completa inadequação das coleções à regra; a 100, em que houve a completa adequação da regra pela coleção. Assim, as cores mais claras representam menores taxas de adequação e as mais fortes maiores taxas de adequação.

Pode-se destacar o uso de vocabulário controlado no elemento de metadado *Class* com menor taxa de adequação em 99% nas coleções dos museus da Abolição; Bandeiras e Missões. Possuindo as demais coleções a adequação em 100%. Destaca-se positivamente também o elemento *Materials and Techniques* com adequação em 93% na coleção do Museu da Abolição e 94% no Museu da Inconfidência, possuindo as demais coleções taxas superiores a 97%, exceto a coleção do Museu Villa-Lobos com adequação em 0%.

Destaca-se negativamente os elementos *Inscription*, *Location*, *Measurements* e *Physical Description* com a completa inadequação em todas as coleções.

O elemento *Creation Location* apresentou menor taxa de adequação nas coleções dos museus da Abolição, Missões, Arqueologia de Itaipu e de Arte Religiosa e Tradicional com adequação em 0%, devido ao elemento não apresentar o uso de taxonomia nessas coleções. Ainda neste elemento, podemos observar que as coleções dos museus Casa Histórica de Alcântara; Casa da Hera, Histórico Nacional – Coleção Museológica, Regional Casa dos Ottoni, Vitor Meirelles, Villa-Lobos, do Diamante e do ouro apresentaram os usos de taxonomia, porém com objetos com valores vazios. Neste caso sem valor taxonômico informado. Os demais apresentaram o uso de taxonomia e todos os objetos haviam valores informados.

Para o elemento *Creator* pode-se observar valores superiores a 81% com exceção das coleções dos museus Casa Histórica de Alcântara, Museu Histórico Nacional – Moedas de Ouro, Victor Meirelles, Villa-Lobos, da Abolição, de Arqueologia de Itaipu e de Arte Religiosa e Tradicional com taxa de adequação em 0%.

Outro elemento que apresentou alguma taxa de adequação é o *Work Type* com 0% na maioria das coleções, exceto as coleções do Museu Vitor Meirelles com 77% de adequação e 100% para as coleções dos museus Regional do Caeté, Solar Monjardim – Coleção Museológica, e o da Abolição.

5. Considerações Finais

O CCO não é um padrão de metadados, mas os conceitos e elementos apresentados podem ser mapeados para vários elementos de metadados (BACA et al., 2006), como, por exemplo, o *Machine-Readable Cataloging* (MARC21), o *Metadata Object Description Schema* (MODS), o Dublin Core, o VRA Core, e, desta forma, também pode ser mapeado para os elementos descritivos do INBCM, conforme se comprovou no alinhamento (Apêndice A).

Destaca-se ainda que o uso de vocabulário controlado é recomendado por 9 (nove) dos 15 elementos de descrição alinhados entre o CCO e INBCM. Desses 9 (nove), apenas 5 (cinco) deles apresentaram algum índice de adequação nas coleções avaliadas, a saber: *Class*, *Creation Location*, *Creator*, *Materials and Techniques* e *Work Type*. O elemento *Class*, alinhada com o elemento INBCM Classificação, foi a que apresentou melhor índice de adequação dentre as coleções analisadas. Pondera-se, assim, que o uso de taxonomia pelos museus do Ibram, por meio do metadado classificação, é um aspecto importante na qualidade de dados das coleções, pois normaliza e padroniza a terminologia que será usada nos processos de busca e recuperação da informação (LANCASTER, 2004), além de ajudar no alcance da interoperabilidade semântica dos dados entre diferentes esquemas de metadados e aplicações (ZENG, 2019).

Finalmente, o diagnóstico permitiu aferir que apesar desses museus e coleções estarem sob gestão do Ibram, as práticas de

catalogação não estão alinhadas entre elas, com grandes discrepâncias nas práticas utilizadas a exemplo dos elementos discricionais *Work Type* e *Creator*. Podemos aferir ainda que os dados das coleções avaliadas carecem de um tratamento mais adequado nos elementos *Inscription, Location*, *Measurements*, *Physical Description* e *Work Type*. Por outro lado, as coleções se mostraram qualificadas em termos do uso adequado de taxonomias para o elemento classificação. Desta forma, recomenda-se, que práticas de catalogação maduras oriundas de instrumentos de referência sejam incorporadas na modelagem de metadados das bases de dados dos museus sob gestão do Ibram, visando qualificar seus atuais padrões de documentação por meio de instrumentos de organização da informação e orientados para usuários finais de sistemas de informação.

Referências

ABBAS, June. **Structures for organizing knowledge**: exploring taxonomies, ontologies, and other schema. New York: Neal-Schuman Publishers, 2010.

BACA, Murtha; HARPRING, Patricia; LANZI, Elisa; MCRAE, Linda; WHITESIDE, Ann. **Cataloging cultural objects: a guide to describing cultural works and their images**. Chicago: American Library Association, 2006.

BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim. Linked Data - The Story So Far. **International Journal on Semantic Web and Information Systems (IJSWIS)**, v. 5, n. 3, p. 1–22, 2009.

BRASIL. MINISTÉRIO DO TURISMO. INSTITUTO BRASILEIRO DE MUSEUS – Ibram. **Museus Ibram – Instituto Brasileiro de Museus**. Brasilia, 2021. Disponível em: https://antigo.museus.gov.br/museus-ibram/. Acesso em: 28 nov. 2021.

CHAPMAN, Arthur D. Principles of Data Quality. Copenhagen, 2005. DOI: 10.15468/DOC.JRGG-A190. Disponível em: https://www.gbif.org/document/80509. Acesso em: 28 jul. 2022.

COELHO, Abeil. **Qualidade_dados_IBRAM**. 2022. Disponível em: https://github.com/AbeilCoelho/Qualidade_dados_IBRAM. Acesso em: 10 set. 2022.

GILLILAND, Anne J. Setting the Stage. In: BACA, Murta. (ed.). **Introduction to metadata**. 3. ed. Los Angeles: Getty Research Institute, 2016. Disponível em: https://www.getty.edu/publications/intrometadata/setting-the-stage/. Acesso em: 22 jul. 2022.

HARPRING, Patricia. **Metadata Standards Crosswalks**. [S. l.], 2022. Disponível em: https://www.getty.edu/research/publications/electronic_publications/intrometadata/crosswalks.html#endnote1CCO. Acesso em: 17 jul. 2022.

LANCASTER, Frederick Wilfrid. **Indexação e resumos**: teoria e prática. Brasília: Briquet de Lemos, 2004.

MACEDO, Dirceu Flávio; LEMOS, Daniela Lucas da Silva. Dados abertos governamentais: iniciativas e desafios na abertura de dados no Brasil e outras esferas internacionais. **AtoZ: novas práticas em informação e conhecimento**, [S. l.], v. 10, n. 2, p. 14, 2021.

MARTINS, Dalton Lopes et al. Information organization and representation in digital cultural heritage in Brazil: Systematic mapping of information infrastructure in digital collections for data science applications. **Journal of the Association for Information Science and Technology**, [S. l.], p. asi.24650, 2022.

WILKINSON, Mark D. et al. The FAIR guiding principles for scientific data management and stewardship. **Scientific Data**, [S. l.], v. 3, n. 1, p. 160018, 2016.

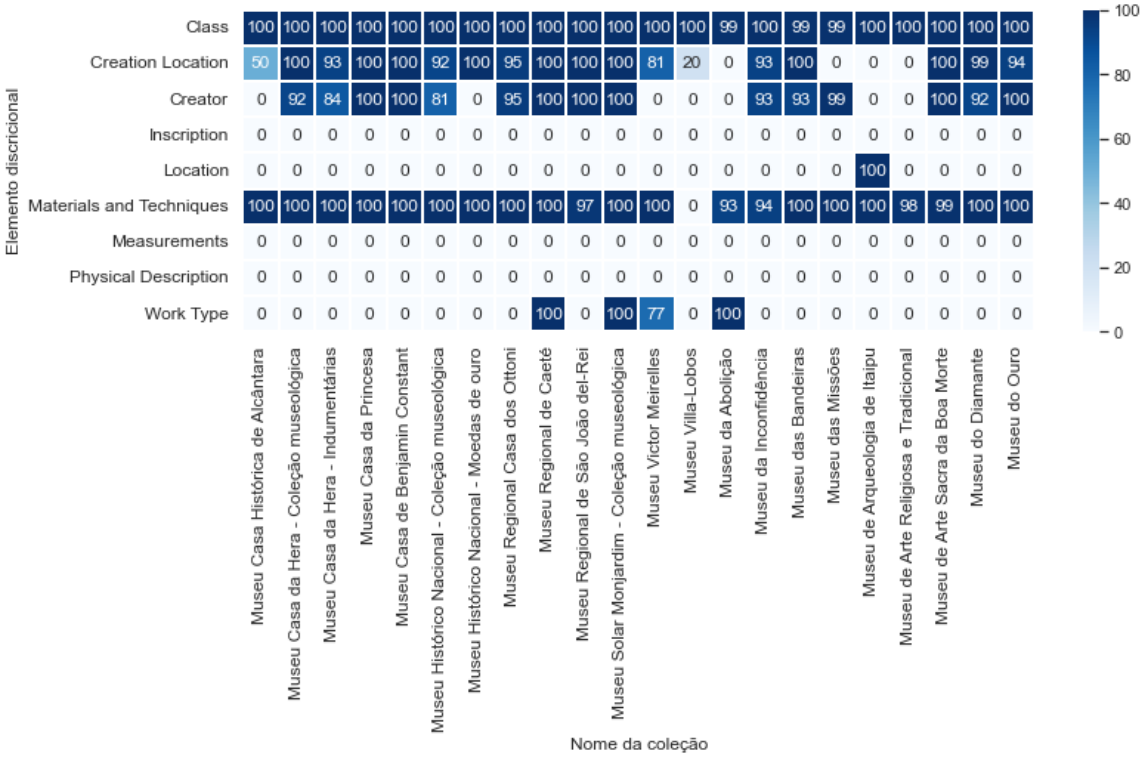
ZENG, Marcia Lei. Interoperability. **Knowledge Organization**, v.46, n.2, p. 122-146, jan. 2019.

Apêndice A – Alinhamento entre elementos descritivos – INBCM e CCO

Capítulo CCO	Elemento CCO	Obrigatório CCO	Vocabulário Controlado CCO	Elemento INBCM	Obrigatório INBCM
I-Part 2	*Work Type*	Sim	Sim	Denominação	Sim
I-Part 2	*Title*	Sim	Não	Título	Não
II-Part 2	*Creator*	Sim	Sim	Autor	Sim
III-Part 2	*Measurements*	Sim	Sim	Dimensões	Sim
III-Part 2	*MaterialsandTechniques*	Sim	Sim	Material/Técnica	Sim
III-Part 2	*Physical Description*	Não	Sim	Estado de Conservação	Sim
III-Part 2	*Inscription*	*Sim*	*Sim*	Número de Registro	Sim
IV-Part 2	*Date*	Sim	Não	Data de Produção	Não
V-Part 2	*CreationLocation*	Não	Sim	Local de Produção	Não
VII-Part 2	*Class*	Sim	Sim	Classificação	Não
VIII-Part 2	*Description*	Não	Não	Resumo Descritivo	Sim
VIII-Part 2	*Other Descriptive Notes*	Não	Não	Condições de Reprodução	Sim
VI-Part 1	*Related Works*	*Não*	*NA*	Mídias Relacionadas	Não
V-Part 2	*Location*	*Sim*	*Sim*	Situação	Sim
NA	*NA*	*NA*	*NA*	Outros Números	Não

Fonte: elaborado pelos autores.

Apêndice B – Adequação de coleções Ibram ao uso de vocabulário controlado



Fonte: elaborado pelos autores.

VISUALIZAÇÃO DE DADOS ABERTOS NO CONTEXTO DA PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO: ANÁLISE BIBLIOMÉTRICA DOS ESTUDOS DEFENDIDOS

VISUALIZATION OF OPEN DATA IN THE CONTEXT OF POSTGRADUATE IN INFORMATION SCIENCE: BIBLIOMETRIC ANALYSIS OF THE DEFENDED STUDIES

Francis Bento Marques[1], Yuri Bento Marques[2], Benildes Coura Moreira dos Santos Maculan[3], Renato Rocha Souza[4]

(1) Programa de Pós-Graduação em Gestão & Organização do Conhecimento - UFMG, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, fbmarques@gmail.com

(2) Programa de Pós-Graduação em Gestão & Organização do Conhecimento - UFMG, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, yuri.marques@ifnmg.edu.br

(3) Programa de Pós-Graduação em Gestão & Organização do Conhecimento - UFMG, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, benildes@gmail.com

(4) Programa de Pós-Graduação em Gestão & Organização do Conhecimento - UFMG, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, rsouzaufmg@gmail.com

Resumo

A visualização de dados abertos vem se tornando um indicativo de credibilidade por parte das instituições nos diversos segmentos sociais, incluindo as instituições de pesquisa. No âmbito desta temática, este estudo teve como objetivo apresentar o panorama da produção acadêmica na Pós-Graduação brasileira na área da Ciência da Informação. Definiu-se como recorte de análise os estudos defendidos nas teses e dissertações nos Programas de Pós-Graduação dessa área, a partir da análise das seguintes facetas: ano, modalidade do programa, grau acadêmico dos cursos, tipos de produção, programas e instituições ofertantes. Trata-se de um estudo descritivo e quantitativo, cujos dados foram coletados por meio de um levantamento bibliométrico realizado no Catálogo de Teses e Dissertações, no âmbito do Plano de Dados Abertos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. Os resultados revelaram um crescimento gradativo entre a produção geral, os programas e os cursos, prevalência das dissertações e uma distribuição heterogênea entre os programas e as instituições, localizadas nas diferentes regiões brasileiras. Concluiu-se que as pesquisas defendidas nos Programas de Pós-Graduação estão visíveis no Plano de Dados Abertos, que a área da Ciência da Informação está em desenvolvimento e que os dados permitem análises transversais que podem beneficiar as políticas educacionais.

Palavras-chave: visualização de dados abertos; programas de Pós-Graduação; Ciência da Informação; plano de dados abertos da Capes.

Abstract

The visualization of open data has become an indicator of credibility on the part of institutions in different social segments, including research institutions. Within the scope of this theme, this study aimed to present the panorama of academic production in Brazilian Graduate Studies in the area of Information Science. The studies defended in the theses and dissertations in the Postgraduate Programs in this area were defined as an analysis cut, based on the analysis of the following facets: year, type of program, academic degree of the courses, types of production, programs and offering institutions . This is a descriptive and quantitative study, whose data were collected through a bibliometric survey carried out in the Catalog of Theses and Dissertations, within the scope of the Open Data Plan of the Coordination for the Improvement of Higher Education Personnel. The results revealed a gradual growth between the general production, programs and courses, prevalence of dissertations and a heterogeneous distribution between programs and institutions, located in different Brazilian regions. It was concluded that the research defended in the Graduate Programs are visible in the Open Data Plan, that the area of Information Science is in development and that the data allow cross-sectional analysis that can benefit educational policies.

Keywords: open data visualization; Graduate programs; Information Science; Capes open data plan.

1 Introdução

A visualização de dados constitui uma estratégia de comunicação e de transparência acerca do que é produzido pelas instituições. No âmbito científico, essa tendência fortaleceu-se nos últimos anos por diversos fatores, como: a necessidade de se preservar a memória institucional, o grande

volume de dados produzidos, a facilidade oportunizada pelas tecnologias e a obrigação em dar transparência às atividades desenvolvidas, sobretudo em instituições do setor público.

Silva (2018) corrobora esse argumento, mencionando que as instituições públicas de pesquisa, ao disponibilizarem os dados produzidos, comunicam à sociedade os resultados de suas ações, com a finalidade de justificar os investimentos e garantir que tais dados sejam reutilizados.

A fim de permitir que os dados produzidos sejam comunicados à sociedade, essas instituições sistematizam esses dados em repositórios, muitas vezes, em ambientes digitais, dando a eles acesso aberto (disponibilizado gratuitamente, sem restrições). Nesses espaços, a sociedade pode acessar os dados e realizar análises e visualizações, que tornam as informações claras e promovem o entendimento sobre "[...] as interconexões e os relacionamentos causa-efeito que estão por trás de cada atividade ou conjunto de políticas públicas" (SILVA, 2018, p. 523).

O acesso aos dados abertos promove a transparência pública, legitima o trabalho das instituições e permite a participação da sociedade, sobretudo com o uso de ferramentas digitais de manipulação e gerenciamento de dados (MACEDO *et al*., 2020). No contexto científico, a disponibilização dos dados de pesquisa pelas instituições de fomento, por exemplo, além de enaltecer o trabalho da instituição, demonstra o desenvolvimento das diversas atividades realizadas pelas áreas do conhecimento.

Uma das agências de fomento que se preocupa com a disponibilização dos dados de pesquisa é a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Essa agência, no ano de 2017, elaborou o Plano de Dados Abertos (PDA), com o intuito de publicizar as suas ações e estratégias, que nortearão as atividades de implementação e promoção da abertura de dados. A priorização dos conjuntos de dados a serem disponibilizados foi definida considerando a relevância das informações para o cidadão e o grau de maturidade dos conjuntos de dados dentro da instituição (CAPES, 2022).

Na visão de Torino, Trevisan e Vidotti (2019), os dados abertos da CAPES representam uma iniciativa de adequação à Lei de Acesso à Informação, Lei n. 12.527, de 18 de novembro de 2011, que determinou a obrigatoriedade dos órgãos públicos em dar transparência de suas atividades aos cidadãos. Além disso, a legislação exigiu às instituições a facilidade de acesso, "[...] no âmbito de suas competências, de informações de interesse coletivo ou geral por eles produzidas ou custodiadas" (BRASIL, 2011, não paginado).

Assim, o PDA pode ser considerado uma importante estratégia para demonstração das atividades, do desempenho, dentre outros elementos que caracterizam uma área do conhecimento. Mesmo apresentando falhas no que tange ao tratamento dos metadados, conforme resultados do estudo de Torino, Trevisan e Vidotti (2019, p. 45), no PDA, "[...] todos os dados estão estruturados e representados para que possam ser facilmente compreendidos pelos consumidores [...]".

Nesse sentido, entende-se que os dados disponibilizados pelas agências possibilitam a medição do desempenho das áreas científicas, por meio de diferentes indicadores como: produtividade, autores envolvidos, instituições, temáticas, períodos, colaborações, dentre outras categorias. Esse processo pode ser concretizado por meio dos estudos bibliométricos que, segundo Wolski *et al*. (2021), constituem o conjunto de técnicas estatísticas que mensuram índices de produção e disseminação do conhecimento científico.

As análises dos conjuntos de dados abertos oportunizam novas verificações e práticas. Para Gonçalves *et al*. (2018), o trabalho com esses dados pode ser realizado por diferentes áreas do conhecimento, tendo em vista identificar o conhecimento útil em grandes volumes de dados. Os autores mencionam que conjuntos específicos de dados poderão proporcionar descobertas valiosas, a partir do desdobramento de novos estudos. Assim, nota-se que a abrangência do PDA vislumbra oportunidades de mapeamento das características de cada área do conhecimento, embora ainda se note, na literatura, falta de estudos que quantifiquem os dados referentes a áreas

específicas, a exemplo da Ciência da Informação.

Nesse contexto, o presente estudo versa sobre a visualização de dados abertos à luz do PDA e com foco na área da Ciência da Informação. O estudo constitui o desdobramento de uma tese sobre visualização de dados e apresenta dados preliminares sobre os Programas de Pós-Graduação e suas instituições no âmbito dessa área do conhecimento.

Justifica-se a escolha desta temática por seu ineditismo e por sua importância. Para a Ciência da Informação, analisar o PDA representa mais uma estratégia de conhecer o desenvolvimento dessa área, sobretudo por ser uma área nova, ainda em processo de consolidação. Pelos resultados, tornar-se-á possível conhecer o panorama das pesquisas dessa área, de modo que as instituições e os pesquisadores fortaleçam os investimentos despendidos e ampliem as práticas de pesquisa.

2 Objetivos

A partir da temática delimitada para este estudo, a proposta direciona-se à representatividade da área da Ciência da Informação, pois investiga-se se a identidade da área por meio da análise ao conjunto de dados do PDA.

O objetivo central do estudo é apresentar o panorama científico da produção acadêmica referente às teses e dissertações na Pós-Graduação brasileira, na área da Ciência da Informação.

3 Procedimentos metodológicos

A metodologia utilizada se caracteriza como descritiva, porque expõe os detalhes sobre o objeto investigado, com foco em suas características e considerando variáveis a ele relacionadas (GIL, 2010). Tem abordagem quantitativa, visto que os dados são analisados em sua objetividade, por meio de técnicas estatísticas que mensuram as variáveis (CRESWEL, 2007).

Quanto aos procedimentos técnicos, aplicou-se um estudo bibliométrico que, neste estudo, se refere a um levantamento de informações sobre uma área do conhecimento, com o propósito de identificar aspectos característicos dessa área. Esse levantamento é realizado considerando vários atributos, o que facilita a identificação de tendências (QUEVEDO-SILVA *et al*., 2016).

Os dados foram coletados no Catálogo de Teses e Dissertações (CTD) que consolidam as informações da Capes sobre a produção de teses e dissertações. Foram utilizados os dados referentes a dois CTD: de 2013 a 2016; e de 2017 a 2020.

No primeiro CTD há dados de teses e dissertações defendidas entre 2013 a 2016 e contém 56 tipos de dados, sendo adicionados códigos de identificação para instituição, discente, docente, data de entrada e saída do programa, o endereço para o texto completo da dissertação ou tese na Plataforma Sucupira, dentre outros detalhes. Por fim, no segundo CTD, os dados referem-se aos anos de 2017 a 2020 e possuem 58 variáveis. Nesse catálogo, foram adicionados mais dois campos em relação ao segundo, que mostram a existência de estudos vinculados a essa produção (como artigos de periódicos, de eventos, dentre outros) e, caso haja, qual é a identificação em outra base de dados (CAPES, 2022).

A partir da identificação e caracterização desses catálogos, delimitou-se o recorte de análise. Assim, foram selecionados os dados do código “60700009”, referentes à Ciência da Informação. Nesse processo, os dados foram extraídos em correlação com às oito variáveis ou categorias a serem mapeadas e analisadas, a saber: 1) ano de defesa; 2) modalidade de programa; 3) modalidade de programa por no; 4) grau acadêmico do curso; 5) grau acadêmico por anos; 6) tipos de produção; 7) programas; e 8) instituição.

A coleta de dados foi realizada de modo manual, com *download* de 8 arquivos em formato CSV. Devido ao tamanho dos arquivos, foi utilizado o programa Openrefine para abrir os mesmos e filtrar os trabalhos da área do conhecimento “60700009”. Após o filtro aplicado, foi criado um arquivo com teses e dissertações da área selecionada.

Para a análise dos dados, recorreu-se ao uso das técnicas de Estatística Descritiva (para descrever e resumir conjuntos de dados), cujos resultados foram expostos no formato de gráficos e confrontados com estudos correlatos à temática. No tratamento dos dados para visualização gráfica, adotou-

se a linguagem Python e bibliotecas, como Pandas, Seaborn, Matplotlib e Plotly.

4 Resultados

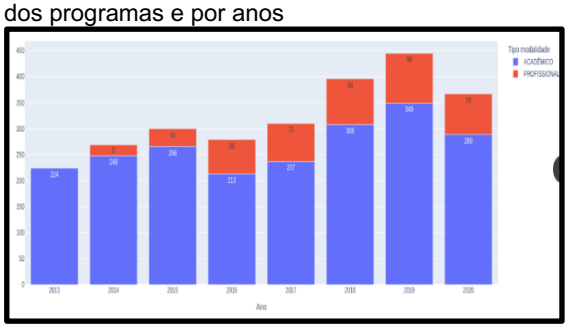
Foram identificados um total de 2590 teses e dissertações no período de 2013 a 2020. Ao distribuir esse quantitativo por ano de defesa, os dados indicaram o ano de 2019 como o mais produtivo (445 estudos), seguido pelo ano de 2018 (396) e 2020 (367). Os demais anos apresentaram defesas oscilante entre 224 (em 2013) e 310 (em 2017).

Esse resultado evidencia, em linhas gerais, um crescimento gradativo de defesas na área. Para Lança *et al*. (2018), essa área vem crescendo como campo de pesquisa em virtude de um aumento da produção dos programas. Assim, "[...] ganha cada vez mais espaço devido às suas contribuições relacionadas à gestão do conhecimento científico e tecnológico [...]" (LANÇA *et al*. 2018, p. 4555). Nota-se decréscimo (em relação ao ano anterior) apenas nos anos: 2016 e 2020. Neste último ano, pode ser devido ao fato que o mundo vivia a dura realidade de uma pandemia.

A segunda categoria diz respeito à modalidade de programa expressa nos dados, que classifica os Programas como Acadêmico ou Profissional. Os resultados identificaram 2134 teses e dissertações na modalidade acadêmica e 456 na profissional, cabendo ressaltar que na Ciência da Informação ainda não há programas com doutorado na modalidade profissional.

Para a terceira categoria, referente a modalidade de programa por anos, foram mapeadas as defesas por anos, conforme mostra o Gráfico 1.

Gráfico 1 – Distribuição dos estudos por modalidades dos programas e por anos

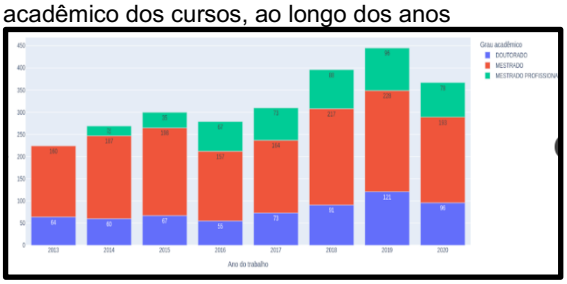


Fonte: Dados da pesquisa (2022).

Percebe-se um crescimento gradativo, similar à distribuição das defesas por ano (sem considerar a divisão por modalidades). Na modalidade acadêmica, houve diminuição em dois anos (2016 e 2020), e apenas em um ano (2020) para a modalidade profissional. Nota-se um crescimento, também gradativo, quando se individualiza por modalidade (acadêmica e profissional), o que contribuiu, de modo proporcional, para o crescimento da área como um todo. O número menor de programas profissionais pode ser pelo fato de que essa modalidade é mais recente, embora sejam bastante adequados à Ciência da Informação, por direcionar-se à resolução de problemas empíricos vivenciados nas instituições (LANÇA *et al*., 2018).

Por sua vez, na quarta e quinta categorias, os estudos foram distribuídos por grau acadêmico do curso e grau acadêmico por anos. Os dados mostram que o Mestrado acadêmico produziu 1504 defesas, o Doutorado acadêmico 627 e o Mestrado Profissional 459, com crescimento gradativo ao longo dos anos. Para o Mestrado e Doutorado acadêmicos, constata-se decréscimo de defesas nos anos 2016 e 2020, e, para o Mestrado Profissional, a diminuição ocorreu, apenas, em 2020, conforme exposto no Gráfico 2.

Gráfico 2 – Distribuição dos estudos por grau acadêmico dos cursos, ao longo dos anos



Fonte: Dados da pesquisa (2022).

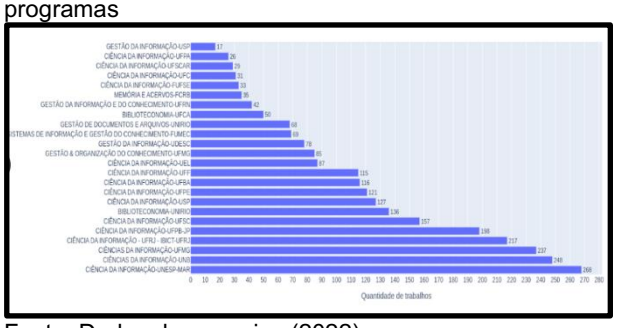
Os dados do Gráfico 2 possibilitam a mesma revelação apreendida pela análise do Gráfico 1, de que os programas e cursos da área se fortaleceram, ao longo dos anos, haja vista o crescimento do número de pesquisas. Assim, nota-se uma relação de proporcionalidade entre os crescimentos gradativos dessas duas variáveis com o crescimento gradativo do número de estudos como um todo, sem considerar as variáveis. Costa, Oliveira e Araújo (2019) também

levantaram um número expressivo de estudos, constatando que há forte impacto acadêmico deles no âmbito da ciência, tanto nacional quanto internacional. Martins *et al*. (2015) destacaram que o aumento dos estudos está relacionado às demandas e oportunidades do mercado, sendo que o Mestrado predomina, sobretudo, por sua tradição como o primeiro nível de Pós-Graduação, como também por ser um curso de curta duração.

Na categoria seis, a distribuição dos estudos por tipos de produção revelou a predominância do tipo dissertação (1907 estudos) em detrimento às teses (627), produto/processo/técnica (48), projeto técnico (6), editoria (1) e relatório final de pesquisa (1). Esse resultado pode ser explicado por existir uma quantidade maior de cursos de Mestrado. Ressalta-se que as teses e dissertações representam o principal produto da Pós-Graduação, pois relatam o estado da arte sobre um determinado assunto, apresentam um rico material metodológico e norteiam a condução de novas descobertas (COSTA; OLIVEIRA; ARAÚJO, 2019).

Para a sétima categoria, verificou-se a distribuição dos estudos por Programas de Pós-Graduação. Foi possível mapear a existência de 24 programas, os quais apresentam diversidade no número de estudos defendidos, conforme ilustrado no Gráfico 3.

Gráfico 3 – Distribuição dos estudos defendidos por programas



Fonte: Dados da pesquisa (2022).

Resultados semelhantes são alcançados quando se analisa a oitava categoria, sobre a distribuição dos estudos por instituições, sendo as três mais produtivas: Universidade Federal de Minas Gerais (322 estudos), Universidade Estadual Paulista (268) e Universidade de Brasília (248), e as três menos produtivas, as universidades: Federal do Pará (26), Federal de São Carlos (29) e Federal do Ceará (31).

Para a sétima e oitava categorias (por programas e por instituições), além da diversidade quanto à quantidade de estudos defendidos, constatou-se que estão presentes em todas as regiões brasileiras. Esse resultado também foi constatado por Moreira e Ferneda (2020), ao relatarem que alguns programas reúnem mais da metade do total de produção científica da área.

5 Conclusão

Com este estudo foi possível apresentar um panorama inicial da produção acadêmica defendida na Pós-Graduação brasileira na área da Ciência da Informação à luz do PDA. Os resultados revelam um crescimento gradativo entre a produção geral, os programas e os cursos; a prevalência das dissertações; e uma distribuição heterogênea entre os programas e as instituições localizadas nas diferentes regiões brasileiras.

Concluiu-se que o PDA possibilita a visualização dos estudos em diferentes áreas do conhecimento. Por meio do PDA foi possível evidenciar que a Ciência da Informação vem se fortalecendo ao longo do tempo. Essa conclusão evidencia o crescente prestígio da área e dá transparência quanto aos investimentos públicos despendidos.

A ausência de estudos correlatos para fortalecer a discussão dos resultados constituiu a limitação da pesquisa. Para pesquisas posteriores, sugere-se analisar as temáticas dos estudos desenvolvidos, como também mapear e comparar mais de uma área do conhecimento.

Referências

BRASIL. **Lei n. 12.527**, de 18 de novembro de 2011. Regula o acesso a informações [...]. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 08 set. 2022.

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR (CAPES). **Plano de dados abertos**. Brasília, set. 2022. Disponível em: https://www.gov.br/capes/pt-br/centrais-de-conteudo/documentos/PlanodeDadosAbertos

daCAPES20202022.pdf. Acesso em: 08 set. 2022.

COSTA, Belkiz; OLIVEIRA, Marlene; ARAÚJO, Ronaldo. Impactos das teses e dissertações do Programa de Pós-Graduação em Ciência da Informação da UFMG. **Informação em Pauta**, Fortaleza, v. 4, n. 2, p. 11-31, jul./dez. 2019. Disponível em: http://www.periodicos.ufc.br/informacaoempauta/article/view/42444/99878. Acesso em: 08 set. 2022.

CRESWEL, John. **Projeto de pesquisa**: método qualitativo, quantitativo e misto. 2. ed. Porto Alegre: Artmed, 2007.

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

GONÇALVES, Alexandre Leopoldo *et al*. Análise de agrupamentos sobre textos. *In*: CONGRESSO INTERNACIONAL DE CONHECIMENTO E INOVAÇÃO, 8., 2018, Guadalajara. **Anais** [...]. Guadalajara: Universidade de Guadalajara, 2018. p. 1-14.Disponível em: https://proceeding.ciki.ufsc.br/index.php/ciki/article/view/589/246. Acesso em: 08 set. 2022.

LANÇA, Tamie Aline *et al*. Produção científica dos Programas de Pós-Graduação em Ciência da Informação na Plataforma Lattes. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 19., 2018, Londrina. **Anais** [...]. Londrina: UEL, 2018. p. 4555-4575.

MACEDO, Daiane *et al*. Uma ferramenta para recomendação de visualização de dados governamentais abertos. *In*: WORKSHOP DE COMPUTAÇÃO APLICADA EM GOVERNO ELETRÔNICO, 8., 2020, Niterói. **Anais** [...]. Niterói: UFF, 2020. p. 1-12. Disponível em: https://sol.sbc.org.br/index.php/wcge/article/view/11261/11124. Acesso em: 07 set. 2022.

MARTINS, Ana Carolina de Melo. Mestrado Profissional na área de Ciência da Informação no Brasil. **Revista ACB**, Florianópolis, v. 20, n. 3, p. 411-422, set./dez. 2015. Disponível em: https://revistaacb.emnuvens.com.br/racb/article/view/1082. Acesso em: 08 set. 2022.

MOREIRA, Jonathan Rosa; FERNEDA, Edberto. Produção científica nos Programas de Pós-Graduação nas áreas de informação no Brasil. **Informação & Informação**, Londrina, v. 25, n. 4, p. 1-20, jul./dez. 2020. Disponível em: https://uel.br/revistas/uel/index.php/informacao/article/view/39670/pdf. Acesso em: 08 set. 2022.

QUEVEDO-SILVA, Filipe *et al*. Estudo bibliométrico: orientações sobre sua aplicação. **Revista Brasileira de Marketing**, São Paulo, v. 15, n. 2, p. 246-262, abr./jun. 2016. Disponível em: https://www.redalyc.org/pdf/4717/471755312008.pdf. Acesso em: 08 set. 2022.

SILVA, Fabiano Couto Corrêa da. Visualização de dados abertos no setor público. **Revista Ibero-americana de Ciência da Informação**, Brasília, v. 11, n. 2, p. 523-540, maio/ago. 2018. Disponível em: https://periodicos.unb.br/index.php/RICI/article/view/8341/9632. Acesso em: 07 set. 2022.

TORINO, Emanuelle; TREVISAN, Gustavo Lunardelli; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Dados abertos CAPES: um olhar à luz dos desafios para publicação de dados na web. **Ciência da Informação**, Brasília, v. 48, n. 3, p. 38-46, set./dez. 2019.Disponível em: https://brapci.inf.br/index.php/res/download/139033. Acesso em: 08 set. 2022.

WOLSKI, Luciano Zamperetti *et al*. Mineração de texto e clusterização em estudos bibliométricos. *In*: CONGRESSO INTERNACIONAL DE CONHECIMENTO E INOVAÇÃO, 11., 2021, Maringá. **Anais** [...]. Maringá: Unicesumar, 2021. p. 1-15.Disponível em: https://proceeding.ciki.ufsc.br/index.php/ciki/article/view/1036. Acesso em: 07 set. 2022.