Ciência da Informação
v. 49 n.3 set./dez. 2020
ISSN 0100-1965 eISSN 1518-8353

Edição especial temática

Special thematic issue / Edición temática especial

Ciência de dados na ciência da informação

Data science in Information Sience

Ciencia de datos en la Ciencia de la Información

Indexação
Ciência da Informação tem seus artigos indexados ou resumidos.

Bases Internacionais
Directory of Open Access Journals - DOAJ. Paschal Thema: Science de L'Information, Documentation. Library and Information Science Abstracts. PAIS Foreign Language Index. Information Science Abstracts. Library and Literature. Páginas de Contenido: Ciências de la Información. EDUCACCION: Notícias de Educación, Ciencia y Cultura Iberoamericanas. Referativnyi Zhurnal: Informatika. ISTA Information Science & Technology Abstracts. LISTA Library, Information Science & Technology Abstracts. SciELO Scientific Electronic Library On-line. Latindex – Sistema Regional de Información em Línea para Revistas Científicas de América Latina el Caribe, España y Portugal, México. INFOBILA: Información Bibliotecológica Latinoamericana.

Indexação em Bases de Dados Nacionais
Portal de Periódicos
LivRe – Portal de Periódicos de Livre Acesso. Comissão Nacional de Energia Nuclear (Cnen). Portal Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes).

Portal de Associações Nacionais
Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação (Ancib).

Bases de Dados Nacionais
Base de Dados Referenciais de Artigos de Periódicos de Ciência da Informação da Universidade Federal do Paraná (Brapci). Escola de Ciência da Informação da Universidade Federal de Minas Gerais (Peri).

Editada em abril de 2021.
Última edição em julho de 2021.
Publicada em julho de 2021.

Ciência da Informação
v. 49 n.3 set./dez. 2020

ISSN 0100-1965 eISSN 1518-8353

Edição especial temática

Special thematic issue / Edición temática especial

Ciência de dados na ciência da informação

Data science in Information Sience

Ciencia de datos en la Ciencia de la Información

Equipe técnica

Editora científica
Cecília Leite Oliveira

Coordenadores editoriais do número
André Luiz Appel
Ricardo Barros Sampaio
Tiago Emmanuel Nunes Braga

Editor executivo
Ramón Martins Sodoma da Fonseca

Editor assistente
Gislaine Russo de Moraes Brito
Alexandre Ribeiro da Silva

Revisão gramatical
Margaret de Palermo Silva
Poliana Martins
Flavia Karla Ribeiro Santos

Diagramação
Dayane Jacob de Oliveira

Projeto gráfico
SEDIT

Capa
Rodrigo Azevedo Moreira

Tradução
SEDIT/Ibict

Normalização de referências
Larissa de Araújo Alves
Elton Mártires Pinto
Ingrid Torres Schiessl
Joyce Mirella dos Anjos Viana

NOTAS DO EDITOR
Para baixar o PDF de cada artigo da revista *Ciência da Informação* a partir do seu smartphone ou tablet, escaneie o QR Code publicado em cada artigo da versão impressa.
Mais informações pelo telefone: (61) 3217-6231

Ciência da Informação/Instituto Brasileiro de Informação em Ciência e Tecnologia
– Vol. 1, n. 1 (1972) – Brasília: Ibict, 1972 –
Quadrimestral

Até o v. 20, 1991, publicada semestralmente. De 1972 a 1975 editada pelo Instituto Brasileiro de Bibliografia e Documentação (IBBD).
ISSN impresso 0100-1965. eISSN 1518-8353.

1. Ciência da Informação – Periódicos I. Brasil, Instituto Brasileiro de Informação em Ciência e Tecnologia.

CDU 02 (05)
CDD 020.5

Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)

Setor de Autarquias Sul (SAUS)
Quadra 05, Lote 06, Bloco H – 5º Andar
Cep: 70070-912 – Brasília, DF
Telefones: 55 (61) 3217-6360
55 (61) 3217-6350
www.ibict.br

Rua Lauro Muller, 455 - 4º Andar - Botafogo
Cep: 22290-160 – Rio de Janeiro, RJ
Telefones: 55 (21) 2275-0321
Fax: 55 (21) 2275-3590
http://www.ibict.br/capacitacao-e-ensino/pos-graduacao-em-ciencia-da-informacao
http://www.ppgci.ufrj.br

Comitê Editorial (março de 2019 a março de 2021)

AVALIADORES DESTE NÚMERO

Cláudio José Silva Ribeiro
Pós-Doutorado pela University of Twente (UT) - Holanda. Doutor em Ciências da Informação pela Universidade Federal Fluminense (UFF) - Brasil. Professor da Universidade Federal do Estado do Rio de Janeiro (UNIRIO) – RJ - Brasil.
http://lattes.cnpq.br/1459853686434404
E-mail: claudio.j.s.ribeiro@globo.com

Eduardo Couto Dalcin
Doutor em Biodiversity Informatics pela University of Southampton (SOUTHAMPTON) - Inglaterra. Coordenador do Núcleo de Computação Científica e Geoprocessamento do Instituto de Pesquisas Jardim Botânico do Rio de Janeiro (JBRJ) - Brasil.
http://lattes.cnpq.br/8334174268306003
E-mail: edalcin@jbrj.org

Fabiano Couto Corrêa da Silva
Doutor em Información y documentación Sociedad Conocimiento pela Universitat de Barcelona (UB) - Espanha. Professor da Universidade Federal do Rio Grande do Sul (UFRGS) – RS - Brasil.
https://orcid.org/0000-0001-5014-8853
http://lattes.cnpq.br/4635807083312321
E-mail: fabianocc@gmail.com

Fernanda Gomes Almeida
Doutora em Gestão & Organização do Conhecimento pela Universidade Federal de Minas Gerais (UFMG) – MG - Brasil. Bibliotecária da Universidade Federal de Minas Gerais (UFMG) - Brasil.
https://orcid.org/0000-0001-7913-827X
http://lattes.cnpq.br/5601300780102290
E-mail: usernanda@gmail.com

Guilherme Ataíde Dias
Pós-Doutorado pela Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp) - Brasil. Doutor em Ciências da Comunicação /Ciência da Informação pela Universidade de São Paulo (USP) – SP - Brasil. Professor da Universidade Federal da Paraíba (UFPB) – PB - Brasil.
https://orcid.org/0000-0001-6576-0017
http://lattes.cnpq.br/9553707435669429
E-mail: guilhermeataide@gmail.com

Maira Murrieta Costa
Doutora em Ciências da Informação pela Universidade de Brasília (UnB) – DF – Brasil, com período sanduíche em University of Michigan - Estados Unidos. Tecnologista do
Ministério da Ciência, Tecnologia e Inovação (MCTI) - Brasília, DF - Brasil
http://lattes.cnpq.br/0580168449333057
https://orcid.org/0000-0002-8324-2114
E-mail: mairamurrieta@gmail.com

Patricia Corrêa Henning
Pós-Doutorado pela University of Twenty (UT) - Holanda. Doutora em Informação e Comunicação em Saúde pelo Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT) – Brasil, com período sanduíche em Universidade de Coimbra (UC) - Portugal. Professora da Universidade Federal do Estado do Rio de Janeiro (Unirio) – RJ - Brasil.
http://lattes.cnpq.br/0970010723997242
E-mail: henningpatricia@gmail.com

Patricia Rocha Bello Bertin
Doutora em Information Management pela Loughborough University (LBORO) - Inglaterra. Pesquisadora e Supervisora de Governança da Informação e Transparência da Secretaria de Desenvolvimento Institucional da Empresa Brasileira de Pesquisa Agropecuária (Embrapa) - Brasil.
http://orcid.org/0000-0001-5973-0305
http://lattes.cnpq.br/4785200171802218
E-mail: patricia.bertin@embrapa.br

Ricardo Barros Sampaio
Pós-Doutorado pela Universidade de Brasília (UnB) – DF – Brasil. Pós-Doutorado pela Fundação Oswaldo Cruz (FIOCRUZ) - Brasil. Doutor em Ciências da Informação pela Universidade de Brasília (UnB) – DF – Brasil, com período sanduíche em Université de Toulouse - França. Professor e pesquisador do Programa de Pós-Graduação em Ciência da Informação da Universidade de Brasília (UnB) – DF – Brasil. Professor e pesquisador no Mestrado Profissional de Políticas Públicas em Saúde e na especialização em Saúde Coletiva pela Escola Fiocruz de Governo - Brasília, DF – Brasil.
http://lattes.cnpq.br/3477515781752110
E-mail: rsampaio.br@gmail.com

Sonia Elisa Caregnato
Doutora em Information Studies pela University of Sheffield (SHEFFIELD) - Inglaterra. Professora da Universidade Federal do Rio Grande do Sul (UFRGS) – RS - Brasil.
https://orcid.org/0000-0002-5676-2763
http://lattes.cnpq.br/5627209208288722
E-mail: sonia.caregnato@ufrgs.br

Vanessa de Arruda Jorge
Pós-Doutorado pela Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) - Brasil. Doutora em Ciência da Informação pela Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) - Brasil. Tecnologista em Saúde Pública da Fundação Oswaldo Cruz (Fiocruz) - Brasil.
https://orcid.org/0000-0002-5298-9311
http://lattes.cnpq.br/0218139692140149
E-mail: vanessaajorge@gmail.com

Viviane Santos de Oliveira Veiga
Doutora em Informação e Comunicação em Saúde pela Fundação Oswaldo Cruz (Fiocruz) – Brasil, com período sanduíche em Universidade de Coimbra (UC) - Portugal. Tecnologista em Saúde Pública da Fundação Oswaldo Cruz (Fiocruz) – Brasil.
https://orcid.org/0000-0001-8318-7912
http://lattes.cnpq.br/4983074089687751
E-mail: vivianesantosveiga@gmail.com

Wagner Junqueira de Araújo
Pós-Doutorado pela Universidade Federal de Pernambuco (UFPE) - Brasil. Doutor em Ciência da Informação pela Universidade de Brasília (UnB) – Brasília, DF - Brasil. Professor da Universidade Federal da Paraíba (UFPB) – PB - Brasil.
https://orcid.org/0000-0002-2301-4996
http://lattes.cnpq.br/6762905361803183
E-mail: wagnerjunqueira.araujo@gmail.com

Ciência da Informação

Volume 49 - número 3 - set./dez. 2020

Sumário

Table of Contents / Sumário

EDITORIAL

O compartilhamento de questões e problemas de pesquisa extrapolam as fronteiras de disciplinas científicas e é o cerne da criação de redes e comunidades entre cientistas, especialistas e organizações. Ao se compartilhar as questões e problemas relacionados a determinada temática, também se socializa as soluções e caminhos que estão sendo tomados por estes grupos que atuam de forma colaborativa, possibilitando abordagens interdisciplinares.

A própria área da Ciência da Informação é interdisciplinar por natureza. Sendo assim, é uma área que está em constante contato com outras áreas do conhecimento e, por isso, é capaz de incorporar elementos até então estranhos à área. A interdisciplinaridade se baseia na articulação de diferentes disciplinas e as coloca em inter-relação. Neste número especial, o foco dos trabalhos se dá justamente à emergente discussão sobre a Ciência de Dados no âmbito de diversas disciplinas e à necessidade de pautar essa discussão de forma mais consolidada na área da Ciência da Informação.

A Ciência de Dados tem como matéria prima o dado, que também é uma das matérias primas utilizadas pela Ciência da Informação. Logo, nada mais natural que a colaboração entre estas duas disciplinas. Um levantamento realizado nas tradicionais bases *Information Science & Technology Abstracts* e *Library, Information Science & Technology Abstracts* mostra que essa tendência de interação com a Ciência de Dados já é uma realidade para a área da Ciência da Informação. Ao se buscar pelo termo "*data science*" percebe-se um crescimento exponencial no número de artigos publicados e que incorporaram a temática. O primeiro artigo a abordar o termo remonta ao ano de 1977, mas é a partir do ano de 2016 que se percebe um maior interesse dos pesquisadores da Ciência da Informação sobre o tema. Naquele ano foram mapeados 148 artigos, o dobro que no ano anterior. Em 2020 foram 281 artigos identificados sobre a temática "*data science*", consolidando um crescimento contínuo que já dura mais de uma década.

No entanto, há de se ressaltar que a Ciência de Dados se refere a um conceito diverso que engloba outros conceitos igualmente abrangentes como *big data*, *machine learning*, *information retrieval*, dentre outros. Logo, o que se pode esperar da interação entre a Ciência de Dados e a Ciência da Informação é justamente a qualificação desses conceitos a partir da perspectiva própria da nossa área. Os artigos presentes neste número primaram pela diversidade.

A revista aborda em seu artigo de abertura aspectos relacionados à relação entre Ciência de Dados e Ciência da Informação com o título: *Interlocuções bibliográficas e epistemológicas entre a ciência de dados e a ciência da informação*. Depois são abordados temas que tratam de como a Ciência de Dados pode ser utilizada para aprimorar o processo de organização e classificação de dados, metadados e a informação gerida por sistemas informacionais. É o caso dos artigos: *Grau de pertencimento como insumo para classificação automática de textos: uma abordagem sintática*, *Dados e metadados*, *Modelagem de metadados multimídia*, *Recuperação de informação: descoberta e análise de workflows para agregação de dados do patrimônio cultural*, *Dicionário Semântico de Dados: abordagem de anotação de dados aplicada à geração de indicadores de desempenho*, *Modelo de Análise Temporal em Contexto Semântico de Gerenciamento de Emergências e Fusão de dados para análise de imagens registradas por satélites: proposta de modelo de metadados*.

Como não podia deixar de ser, há um grande avanço na discussão acerca dos dados abertos, seus repositórios e ferramentas que favorecem a abertura de dados. Fazem parte desse bloco os seguintes artigos: *Explorando a Reconciliação de Dados Culturais na Wikidata*, *Publicando dados de pesquisa*, *DBacademic: Conectando os dados abertos das instituições de ensino do Brasil*, *GOOGLE DATASET SEARCH: Visão geral e perspectivas para indexação e disponibilização de conjuntos de dados científicos abertos*, *Aplicação de Dados Governamentais Abertos à Luz da Ciência da Informação* e *A publicidade de dados abertos pelo tribunal superior eleitoral: o caso do Repositório de Dados Eleitorais*.

Por fim, o último bloco de artigos deste número da revista Ciência da Informação foca nas métricas informacionais e na utilização destas métricas para entender o avanço da pesquisa científica no Brasil.

EDITORIAL

Os artigos que compõem este último bloco são: *Medição da informação científica na Web 2.0*, *Acervos Culturais Brasileiros no Repositório Wikimedia Commons*, *Perfil das orientações e produções das mulheres fundamentado em dados da Plataforma Lattes*, *Um modelo populacional para análise de genealogia acadêmica: Evidências sobre crescimento acadêmico no Brasil* e *Estimando Futuras Colaborações em Dados sobre Atividades Científicas*.

Após a apresentação deste número é possível perceber que há grande sinergia entre as áreas de Ciência de Dados e Ciência da Informação. É preciso aprofundar a compreensão sobre como essas áreas podem se aprimorar e se alimentar a fim de permitir que a informação seja cada vez mais entendida como matéria prima da ciência. Acreditamos que os artigos aqui apresentados contribuirão significativamente para esse processo.

Boa leitura!

Andre Luiz Appel
Doutor em Ciência da Informação pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) e Universidade Federal do Rio de Janeiro (PPGCI-IBICT/UFRJ) – Rio de Janeiro, RJ – Brasil. Bolsista pesquisador do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) – Brasília, DF – Brasil. Pesquisador do Laboratório Interdisciplinar sobre Informação e Conhecimento (Liinc-UFRJ/IBICT) - Brasil.
E-mail: andreappel@ibict.br

Ricardo Barros Sampaio
Pós-Doutorado pela Universidade de Brasília (UnB) – DF – Brasil. Pós-Doutorado pela Fundação Oswaldo Cruz (FIOCRUZ) - Brasil. Doutor em Ciências da Informação pela Universidade de Brasília (UnB) – DF – Brasil, com período sanduíche em Université de Toulouse - França. Professor e pesquisador do Programa de Pós-Graduação em Ciência da Informação da Universidade de Brasília (UnB) – DF – Brasil. Professor e pesquisador no Mestrado Profissional de Políticas Públicas em Saúde e na especialização em Saúde Coletiva pela Escola Fiocruz de Governo - Brasília, DF – Brasil.
http://lattes.cnpq.br/3477515781752110
E-mail: rsampaio.br@gmail.com

Tiago Emmanuel Nunes Braga
Doutor em Ciência da Informação pela Universidade de Brasília (UnB) – Brasília, DF - Brasil. Coordenador Geral de Tecnologias da Informação e Informática (CGTI) e pesquisador do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) - Brasília, DF – Brasil.
E-mail: tiagobraga@ibict.br

EDITORIAL

Sharing research issues and problems crosses the boundaries of scientific fields and is at the heart of creating networks and communities among scientists, experts and organizations. By sharing issues and problems related to a particular theme, the solutions and paths being taken by these groups working collaboratively are also socialized, enabling interdisciplinary approaches.

The field of Information Science is interdisciplinary by nature. As such, it is in constant contact with other fields of knowledge and, therefore, capable of incorporating elements that were previously foreign to the field. Interdisciplinarity is based on the articulation of different disciplines and places them in interrelationship. In this special issue, the focus of the manuscripts is precisely the emerging discussion about Data Science in the scope of several disciplines and the need to guide this discussion in a more consolidated way in the field of Information Science.

Data is Data Science's raw material, which is also one of the raw materials used by Information Science. Therefore, nothing more natural than the collaboration between these two disciplines. A survey carried out in the traditional databases *Information Science & Technology Abstracts* and *Library, Information Science & Technology Abstracts* shows that this trend of interaction with Data Science is already a reality for the field of Information Science. When searching for the term "data science", an exponential growth in the number of articles published that incorporated the theme can be found. The first article to address the term dates back to 1977, but it is from the year of 2016 onwards that Information Science researchers have become more interested in the topic. In that year, 148 articles were mapped, twice as much as in the previous year. In 2020, 281 articles were identified on the theme "data science", consolidating a continuous growth that has lasted more than a decade.

However, it should be noted that Data Science refers to a diverse concept that encompasses other equally comprehensive concepts such as big data, machine learning, information retrieval, among others. Therefore, what can be expected from the interaction between Data Science and Information Science is precisely the qualification of these concepts from the perspective of our field. The articles in this issue are distinguished by their diversity.

In its opening article, the journal addresses aspects related to the relationship between Data Science and Information Science with the title "Bibliographic and epistemological interlocutions between data science and information science". Then, topics that deal with how Data Science can be used to improve the process of organizing and classifying data, metadata and the information managed by informational systems are discussed. This is the case of the articles "Degree of belonging as an input for automatic text classification: a syntactic approach", "Data and metadata", "Multimedia metadata modeling", "Information retrieval: discovery and analysis of workflows for aggregation of cultural heritage data", "Semantic Dictionary of Data: data annotation approach applied to the generation of performance indicators", "Temporal Analysis Model in Semantic Context of Emergency Management" and "Data Fusion for analysis of images recorded by satellites: proposal of a metadata model".

As it should be, there is a great advance in the discussion about open data, its repositories and tools that favor the opening of data. The following articles are part of this block: Exploring the Reconciliation of Cultural Data on Wikidata, Publishing research data, DBacademic: Connecting open data from educational institutions in Brazil, GOOGLE DATASET SEARCH: Overview and perspectives for indexing and making available sets of Open Scientific Data, Application of Open Government Data in the Light of Information Science and Publicity of Open Data by the Superior Electoral Court: the Electoral Data Repository case.

Finally, the last block of articles in this issue of journal *Ciência da Informação* focuses on informational metrics and the use of these metrics to understand the advancement of scientific research in Brazil. The articles that make up this last block are: Measuring scientific information on Web 2.0, Brazilian Cultural Collections in the Wikimedia Commons Repository, Profile of women's orientations and productions based on data from the Lattes Platform, A population model for academic genealogy analysis: Evidence on academic growth in Brazil and Estimating Future Collaboration on Scientific Activities Data.

EDITORIAL

After the presentation of this issue, it is possible to see that there is great synergy between the fields of Data Science and Information Science. It is necessary to deepen the understanding of how they can improve and feed from each other in order to allow information to be increasingly understood as the raw material of science. We believe that the articles presented here will significantly contribute to this process.

Good reading!

André Luiz Appel
PhD in Information Science from the Brazilian Institute of Information in Science and Technology (Ibict) and Federal University of Rio de Janeiro (PPGCI-IBICT/UFRJ) – Rio de Janeiro, RJ – Brazil. Research fellow at the Brazilian Institute of Information in Science and Technology (Ibict) – Brasília, DF – Brazil. Researcher at the Interdisciplinary Laboratory on Information and Knowledge (Liinc-UFRJ/IBICT) - Brazil.
E-mail: andreappel@ibict.br

Ricardo Barros Sampaio
Post-Doctorate at the University of Brasília (UnB) – DF – Brazil. Post-Doctorate from the Oswaldo Cruz Foundation (FIOCRUZ) - Brazil. PhD in Information Sciences from the University of Brasília (UnB) – DF – Brazil, with a sandwich period in University of Toulouse - France. Professor and researcher at the Graduate Program in Information Science at the University of Brasília (UnB) – DF – Brazil. Professor and researcher at the Professional Master's Degree in Public Health Policy and specialization in Public Health at Fiocruz School of Government - Brasília, DF – Brazil.
http://lattes.cnpq.br/3477515781752110
E-mail: rsampaio.br@gmail.com

Tiago Emmanuel Nunes Braga
PhD in Information Science from the University of Brasília (UnB) – Brasília, DF - Brazil. General Coordinator of Information Technologies and Informatics (CGTI) and researcher at the Brazilian Institute of Information in Science and Technology (Ibict) - Brasília, DF – Brazil.
E-mail: tiagobraga@ibict.br

EDITORIAL

Compartir cuestiones y problemas de investigación traspasa los límites de las disciplinas científicas y está en el centro de la creación de redes y comunidades entre científicos, expertos y organizaciones. Al compartir las cuestiones y problemas relacionados con un tema en particular, las soluciones y caminos que están tomando estos grupos que trabajan en colaboración también se socializan, lo que permite enfoques interdisciplinarios.

El área de las Ciencias de la Información en sí es de naturaleza interdisciplinaria. De esta forma, es un área que está en constante contacto con otras áreas del conocimiento y, por lo tanto, es capaz de incorporar elementos que antes eran ajenos al área. La interdisciplinariedad se basa en la articulación de diferentes disciplinas y las coloca en interrelación. En este número especial, el foco de los artículos es precisamente la discusión emergente sobre *Data Science* en el ámbito de varias disciplinas y la necesidad de orientar esta discusión de una manera más consolidada en el área de las Ciencias de la Información.

La ciencia de datos tiene como materia prima los datos, que también es una de las materias primas que utiliza la ciencia de la información. Por tanto, nada más natural que la colaboración entre estas dos disciplinas. Una encuesta realizada en las bases de datos tradicionales *Information Science & Technology Abstracts* y *Library, Information Science & Technology Abstracts* muestra que esta tendencia de interacción con Data Science ya es una realidad para el área de Ciencias de la Información. Al buscar el término "data science", se puede ver un crecimiento exponencial en el número de artículos publicados que incorporaron el tema. El primer artículo que aborda el término remonta a 1977, pero es a partir del año 2016 cuando los investigadores de Ciencias de la Información se han interesado más por el asunto. En ese año se mapearon 148 artículos, el doble que en el año anterior. En 2020 se identificaron 281 artículos sobre el tema "data science", consolidando un crecimiento continuo que ha durado más de una década.

Sin embargo, cabe destacar que *Data Science* se refiere a un concepto diverso que engloba otros conceptos igualmente completos como *big data*, *machine learning*, recuperación de información, entre otros. De esa forma, lo que se puede esperar de la interacción entre Data Science y Ciencias de la Información es precisamente la calificación de estos conceptos desde la perspectiva de nuestro campo. Los artículos de este número se distinguen por su diversidad.

En su artículo de apertura, la revista aborda aspectos relacionados con la relación entre la ciencia de datos y la ciencia de la información con el título: Interlocuciones bibliográficas y epistemológicas entre ciencia de datos y ciencia de la información. Luego, se discuten temas que tratan sobre cómo la ciencia de datos se puede utilizar para mejorar el proceso de organización y clasificación de datos, metadatos e información administrada por sistemas de información. Este es el caso de los artículos: Grado de pertenencia como insumo para la clasificación automática de textos: un enfoque sintáctico, Datos y metadatos, Modelado de metadatos multimedia, Recuperación de información: descubrimiento y análisis de flujos de trabajo para la agregación de datos del patrimonio cultural, Diccionario semántico de datos : enfoque de anotación de datos aplicado a la generación de indicadores de desempeño, Modelo de Análisis Temporal en Contexto Semántico de Manejo de Emergencias y Fusión de Datos para análisis de imágenes registradas por satélites: propuesta de un modelo de metadatos.

Como debe ser, hay un gran avance en la discusión sobre datos abiertos, sus repositorios y herramientas que favorecen la apertura de datos. Los siguientes artículos forman parte de este bloque: Explorando la reconciliación de datos culturales en Wikidata, Publicación de datos de investigación, DBacademic: Conectando datos abiertos de instituciones educativas en Brasil, GOOGLE DATASET SEARCH: Visión general y perspectivas para indexar y poner a disposición conjuntos de Datos científicos abiertos , Aplicación de Datos de Gobierno Abierto a la Luz de la Ciencia de la Información y Publicidad de Datos Abiertos por el Tribunal Superior Electoral: el caso del Repositorio de Datos Electorales.

Finalmente, el último bloque de artículos de este número de la revista *Ciência da Informação* se centra en las métricas de información y el uso de estas métricas para comprender el avance de la investigación científica en Brasil.

EDITORIAL

Los artículos que componen este último bloque son: Medición de información científica en Web 2.0, Colecciones Culturales Brasileñas en el Repositorio de Wikimedia Commons, Perfil de orientaciones y producciones de mujeres con base en datos de la Plataforma Lattes, Un modelo poblacional para análisis de genealogía académica: Evidencia sobre crecimiento académico en Brasil y Estimación de la colaboración futura en datos de actividades científicas.

Tras la presentación de este número, se puede constatar que existe una gran sinergia entre las áreas de Data Science y Information Science. Es necesario profundizar en la comprensión de cómo estas áreas pueden mejorar y alimentarse para permitir que la información se entienda cada vez más como la materia prima de la ciencia. Creemos que los artículos aquí presentados contribuirán significativamente a este proceso.

¡Buena lectura!

André Luiz Appel
Doctor en Ciencias de la Información por el Instituto Brasileño de Información en Ciencia y Tecnología (Ibict) y Universidad Federal de Rio de Janeiro (PPGCI-IBICT / UFRJ) - Rio de Janeiro, RJ - Brasil. Compañero de investigación en el Instituto Brasileño de Información en Ciencia y Tecnología (Ibict) - Brasilia, DF - Brasil. Investigador del Laboratorio Interdisciplinario de Información y Conocimiento (Liinc-UFRJ / IBICT) - Brasil.
Correo electrónico: andreappel@ibict.br

Ricardo Barros Sampaio
Postdoctorado en la Universidad de Brasilia (UnB) - DF - Brasil. Postdoctorado de la Fundación Oswaldo Cruz (FIOCRUZ) - Brasil. Doctorado en Ciencias de la Información por la Universidad de Brasilia (UnB) - DF - Brasil, con un período sándwich en Universidad de Toulouse - Francia.
Profesor e investigador del Programa de Posgrado en Ciencias de la Información de la Universidad de Brasilia (UnB) - DF - Brasil. Profesor e investigador de la Maestría Profesional en Políticas de Salud Pública y especialización en Salud Pública de la Escuela de Gobierno Fiocruz - Brasilia, DF - Brasil.
http://lattes.cnpq.br/3477515781752110
Correo electrónico: rsampaio.br@gmail.com

Tiago Emmanuel Nunes Braga
Doctorado en Ciencias de la Información por la Universidad de Brasília (UnB) - Brasília, DF - Brasil. Coordinador General de Tecnologías de la Información e Informática (CGTI) e investigador del Instituto Brasileño de Información en Ciencia y Tecnología (Ibict) - Brasilia, DF - Brasil.
Correo electrónico: tiagobraga@ibict.br

Artigos

Articles / Artículos

Grau de pertencimento como insumo para classificação automática de textos: uma abordagem sintática

André Fabiano Dyck
Doutorando em Ciência da Informação pela Universidade Federal de Santa Catarina (UFSC) - Florianópolis, SC - Brasil. Mestre em Ciências da Computação pela Universidade Federal de Santa Catarina (UFSC) - Brasil. Analista de Tecnologia da Informação da Universidade Federal de Santa Catarina (UFSC) - Brasil.
http://lattes.cnpq.br/7745380984531130
E-mail: andre.dyck@ufsc.br

Rogério de Aquino Silva
Mestrando em Ciência da Informação pela Universidade Federal de Santa Catarina (UFSC) - Florianópolis, SC - Brasil. Especialização em Business Intelligence pela Instituto Brasileiro de Tecnologia Avançada (IBTA) - Brasil. Cientista de dados do Instituto de Previdência do Estado de Santa Catarina (IPREV) - Florianópolis, SC - Brasil.
http://lattes.cnpq.br/2735959956037192
E-mail: rogerio.aquino@posgrad.ufsc.br

Moisés Lima Dutra
Doutor em Ciências da Computação pela Université Claude Bernarde Lyon 1 (LYON I) - França, com período co-tutela em Universidade Nova de Lisboa (UNL) – Portugal. Professor da Universidade Federal de Santa Catarina (UFSC) - Florianópolis, SC - Brasil.
http://lattes.cnpq.br/1973469817655034
E-mail: moises.dutra@ufsc.br

Gustavo Medeiros de Araújo
Doutor em Engenharia de Automação e Sistemas pela Universidade Federal de Santa Catarina (UFSC) - Florianópolis, SC – Brasil, com período sanduíche em Otto-von-Guericke-Universität Magdeburg – Alemanha. Professor da Universidade Federal de Santa Catarina (UFSC) - Florianópolis, SC - Brasil.
http://lattes.cnpq.br/2609254559240670
E-mail: gustavo.araujo@ufsc.br

Submetido em: 13/09/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021.

RESUMO

Agrupar documentos em categorias é uma das soluções adotadas para agilizar o processo de recuperação de informação, cada vez mais relevante devido à grande oferta de informação existente nos dias atuais. A localização manual de documentos de determinada temática, disponíveis em repositórios digitais, passa pela leitura de título, resumo e palavras-chave, além de posterior avaliação mais detalhada com o intuito de identificar se a publicação pertence ao eixo temático desejado. Considerando o número de publicações existentes num repositório digital, a localização manual de todos os textos desejados de uma determinada temática pode ser trabalhosa e demorada. Esta pesquisa propõe uma técnica para classificação automática de textos que se baseia em questões sintáticas, ou seja, empreende uma comparação de n-gramas, que são combinações de *n-uplas* de palavras identificadas ao longo do texto. Realizou-se uma pesquisa aplicada, de cunho exploratório, que aplicou um tipo de aprendizagem supervisionada, baseada fundamentalmente no modelo de representação dos documentos chamado saco de palavras (*bag-of-words* - BoW). Seu objetivo-macro foi o de classificar textos de maneira geral, de acordo com categorias pré-definidas, por meio da geração e comparação de graus de pertencimento entre os textos, como um dos critérios-chave. Os resultados destas comparações, a partir da utilização de n-grama = 3, demonstram que, na utilização de classificações por n-gramas, quanto maior o número de gramas, e com a retirada das *stop words*, obtém-se um grau de pertencimento reduzido, demonstrando um rigor maior para identificar a combinação (*match*) durante a classificação. Para termos maior confiança nos resultados, é necessário um *corpus* de treinamento maior, para ampliar o número de palavras que caracterizem as categorias pré-definidas, a serem utilizadas na classificação dos textos.

Palavras-chave: Grau de pertencimento. Classificação textual. *Bag-of-Words*. N-Gramas. Ciência da Informação.

Degree of belonging as an input for automatic text classification: a syntactic approach

Abstract

Grouping documents into categories is one of the solutions adopted to streamline the information retrieval process, which is increasingly relevant due to the large amount of information available today. The manual localization of documents of a specific theme, available in digital repositories, involves reading the title, abstract and keywords, in addition to further detailed evaluation in order to identify whether the publication belongs to the desired thematic axis. Considering the number of publications in a digital repository, manually locating all the desired texts on a given topic can be laborious and time-consuming. This research proposes an architecture for automatic classification of texts that is based on syntactic questions, that is, it undertakes a comparison of n-grams, which are combinations of n-pairs of words that are identified throughout the text. An exploratory applied research was carried out, which applied a type of supervised learning, fundamentally based on the document representation model called bag-of-words (BoW). The paper's macro objective was to classify texts in general, according to pre-defined categories, by generating and comparing degrees of belonging between texts, as one of the key criteria. The results of these comparisons, using n-gram = 3, demonstrate that in the use of classifications by n-grams, the greater the number of grams, and with the removal of the stop words, we obtain a reduced degree of belonging, demonstrating greater rigor in identifying the match during the classification. In order to have greater confidence in the results, a larger training corpus is necessary to expand the number of words that characterize the pre-defined categories, to be used in the classification of the texts.

***Keywords**: Degree of belonging. Text Classification. Bag-of-Words. N-Grams. Information Science.*

Grado de pertenencia como entrada para la clasificación automática de texto: un enfoque sintáctico

RESUMEN

La agrupación de documentos en categorías es una de las soluciones adoptadas para agilizar el proceso de recuperación de información, que es cada vez más relevante debido a la gran cantidad de información disponible en la actualidad. La localización manual de documentos de un tema específico, disponibles en repositorios digitales, implica la lectura del título, resumen y palabras clave, además de una evaluación más detallada con el fin de identificar si la publicación pertenece al eje temático deseado. Teniendo en cuenta la cantidad de publicaciones en un repositorio digital, ubicar manualmente todos los textos deseados sobre un tema determinado puede resultar laborioso y llevar mucho tiempo. Esta investigación propone una arquitectura de clasificación automática de textos que se basa en preguntas sintácticas, es decir, realiza una comparación de n-gramos, que son combinaciones de n-pares de palabras que se identifican a lo largo del texto. Se realizó una investigación aplicada de carácter exploratorio, que aplicó un tipo de aprendizaje supervisado, basado fundamentalmente en el modelo de representación de documentos denominado bolsa de palabras (bag-of-words - BoW). Su macro objetivo era clasificar los textos en general, según categorías predefinidas, generando y comparando grados de pertenencia entre textos, como uno de los criterios clave. Los resultados de estas comparaciones, utilizando n-gramo = 3, demuestran que en el uso de clasificaciones por n-gramos, a mayor número de gramos, y con la eliminación de las palabras vacías, obtenemos un grado de pertenencia reducido, demostrando mayor rigor en la identificación del partido durante la clasificación. Para tener una mayor confianza en los resultados, es necesario un corpus de formación más amplio para ampliar el número de palabras que caracterizan las categorías predefinidas, para ser utilizadas en la clasificación de los textos.

***Palabras clave**: Grado de pertenencia. Classificación textual. Bag-of-Words. N-Gramos. Ciencias de la información.*

INTRODUÇÃO

Agrupar documentos em categorias é uma das soluções adotadas para agilizar o processo de recuperação de informação, cada vez mais relevante devido à enorme oferta de informação dos dias atuais. Estas categorias, ou rótulos, podem ser geradas por meio de intervenção humana, geralmente associando-se semântica, que facilitaria a recuperação, ou usando apenas algoritmos computadorizados que utilizam outras características dos textos para agrupá-los, num processo que é um tipo de classificação.

A classificação é uma capacidade inerente do ser humano, que utiliza categorias como ferramentas para entender o mundo, e este processo envolve uma série de etapas. Segundo Piaget, no construtivismo, o sujeito aprende com base na assimilação, na integração e na reorganização de estruturas que lhe permitem interpretar o mundo e interagir com ele. Ainda longe de mapear e simular este processo complexo, a classificação de texto apenas atua na organização de informação por meio de atribuição de rótulos. Em um sentido computacional, classificar é atribuir rótulos aos dados, que, no caso da classificação textual, são as palavras de um documento. A categorização, por outro lado, trata de agrupar documentos semelhantes, não rotulados, com base em alguma medida de similaridade (INGERSOLL; MORTON, 2013). Neste trabalho, concordamos com a distinção feita por Ingersoll e Morton (2013), de que a classificação textual (*text classification*) e a categorização textual (*text clustering*) são visões diferentes sobre os dados. Enquanto a primeira distingue a forma pela qual um dado pertence a uma categoria e não a outra, de modo absoluto, a segunda considera a semelhança entre os dados dentro de uma mesma categoria, atribuindo níveis de especialização.

A oferta de repositórios de documentos, nos quais podemos fazer pesquisas livres para encontrar os mais variados temas, está aumentando. A localização manual de documentos de determinada temática, disponíveis em repositórios digitais, passa pela leitura do título, do resumo e das palavras-chave e posterior avaliação mais minuciosa para identificar se esta publicação é do eixo temático[1] desejado. Considerando o número de publicações existentes num repositório digital, a localização manual de todos os textos desejados de determinada temática pode ser trabalhosa e demorada. O cenário de aplicação desta pesquisa parte do pressuposto de que os eixos temáticos de pesquisa dos Programas de Pós-Graduação (PPG) em Ciência da Informação (CI) possuem um alinhamento com os Grupos de Trabalho (GT) da Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação (ANCIB). Assim, a realização desta pesquisa abre questões como: Quais são as temáticas mais trabalhadas pelos PPG em CI? Qual é o alinhamento das teses e dissertações da área da CI com os temas dos GTs da Ancib? Como localizar teses e dissertações alinhadas com um eixo temático específico relacionado a um GT da Ancib? Este artigo é uma versão revista e ampliada de um trabalho apresentado no III Workshop de Informação, Dados e Tecnologia (WIDaT 2019), que ocorreu em Brasília em novembro de 2019 (DYCK; DUTRA; VIERA, 2019). A estrutura do artigo segue com: (i) os procedimentos metodológicos utilizados; (ii) a análise e discussão dos resultados; e (iii) as considerações finais.

METODOLOGIA

A realização deste trabalho toma por base o modelo de representação dos documentos chamado saco de palavras (*bag-of-words* ou BoW). O BoW é um modelo de representação simplificado usado no processamento de linguagem natural[2], uma subárea da Ciência da Computação, para representar os documentos como um conjunto de palavras, sem considerar sua semântica original (HARRIS, 1954; GOLDBERG, 2017).

Ao trabalhar as coleções de textos como BoW, utilizando apenas as palavras e suas combinações presentes no texto, sem considerar questões semânticas, chamamos de n-gramas essas palavras ou suas combinações. Por exemplo, a palavra "grau", é um exemplo da representação de n-grama=1; as palavras "grau de", são exemplos da representação de n-grama=2; as palavras "grau de pertencimento", são exemplos da representação de n-grama=3; as palavras "grau de pertencimento como", são exemplos da representação de n-grama=4 e; as palavras "grau de pertencimento como insumo", são exemplos da representação de n-grama=5 (MOURA *et. al.*, 2010; JURAFSKY; MARTIN, 2018). As coleções de textos trabalhadas neste artigo como cenário de aplicação se referem: (i) aos resumos e aos textos completos de teses e dissertações do Programa de Pós-Graduação em Ciência da Informação, da Universidade Federal de Santa Catarina (PGCIN/UFSC[3]), extraídos do Repositório Institucional da UFSC – RI/UFSC[4]; e (ii) às ementas dos GTs da Ancib, extraídas do site do "Fórum de Coordenadores de Grupo de Trabalho da Ancib"[5].

No presente estudo, realizamos uma pesquisa aplicada, de cunho exploratório, cuja coleta e tabulação de dados se deram entre os dias 26/08/2019 e 09/09/2019, para os resumos, e entre os dias 12/12/2019 e 21/12/2019, para os textos completos, e que toma como cenário de aplicação o conjunto de 223 documentos das teses e dissertações em CI que estavam disponíveis no RI/UFSC naquele período. Para obtermos os resumos, ao acessar o RI/UFSC, selecionamos a comunidade "Teses e Dissertações" e, então, a coleção "Programa de Pós-Graduação em Ciência da Informação". Em seguida, exportamos os metadados[6] dos 223 documentos para um arquivo CSV que, então, foram lidos pela biblioteca Python Pandas.

[1] Eixo temático no contexto deste trabalho significa um suporte ou guia para limitar os conteúdos de um assunto principal.

[2] Linguagem natural no contexto deste trabalho são as línguas faladas pelos humanos.

[3] http://pgcin.paginas.ufsc.br/

[4] https://repositorio.ufsc.br/

[5] http://gtancib.fci.unb.br/

[6] Metadados, no contexto deste trabalho, são dados sobre as dissertações e teses, como por exemplo: título, resumo, tipo de publicação, palavras-chave etc.

Com o arquivo CSV em memória, extraímos a coluna de resumos. Os textos completos das teses e dissertações dos 223 documentos foram obtidos com a equipe que administra a ferramenta de RI da UFSC, num arquivo compactado em formato ZIP[7].

A redução dos textos às palavras que os constituem forma os unigramas. Uma possibilidade de se preservar um mínimo do significado do texto, usando ainda BoW, é utilizar também bigramas e trigramas (GOLDBERG, 2017). Assim, mantém-se a proximidade de duas e três palavras do texto original. Esta pesquisa identificou e analisou unigramas, bigramas e trigramas, e, também, n-grama = 4 e n-grama = 5, tanto para os resumos quanto para os textos completos de teses e dissertações e para as ementas de GT coletadas, para assim comparar seus resultados.

Este trabalho utilizou técnicas de Aprendizagem de Máquina, uma área de inteligência artificial que está preocupada em desenvolver algoritmos que aprendem padrões presentes em uma massa de dados (chamada de massa de dados de aprendizagem). Estes padrões aprendidos podem ser usados para predizer informações sobre dados novos, por isso a importância da massa de dados de aprendizagem ser diversa o suficiente para ampliar as chances de predições (BAEZA-YATES; RIBEIRO-NETO, 2013).

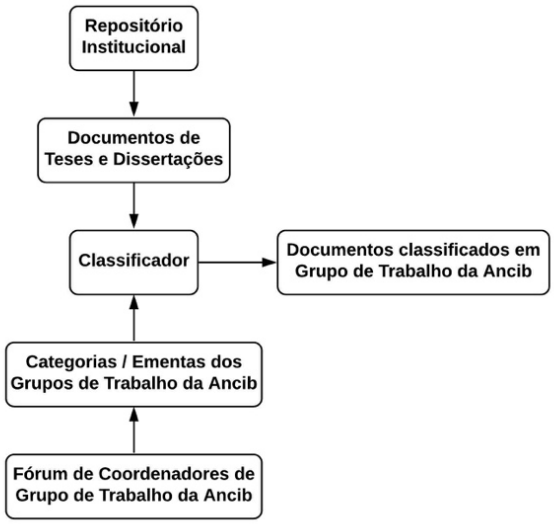
O uso deste tipo de algoritmos é extensivo em diagnóstico médico, detecção de fraudes a cartões de crédito, análise de mercado de ações e recuperação de informação. Na recuperação de informação, a classificação de textos é chave para o sucesso (BAEZA-YATES; RIBEIRO-NETO, 2013). Para automatizar a classificação de texto, podemos fazer uso de várias técnicas e conceitos. Há principalmente três tipos de técnicas de aprendizagem: (i) Aprendizagem de máquina supervisionada, quando há intervenção humana na fase de treinamento; (ii) Aprendizagem de máquina não supervisionada, quando não há intervenção humana no treinamento, como, por exemplo, a técnica chamada de clusterização (*clustering*); e (iii) Aprendizagem semi-supervisionada, na qual o conjunto inicial de dados é composto apenas por uma pequena entrada rotulada e grande parte dos dados da entrada não está rotulada, i.e., a categoria associada a eles é desconhecida. Neste caso, o objetivo é similar à classificação supervisionada, que é gerar uma relação binária, mapeando a entrada para saída (BAEZA-YATES; RIBEIRO-NETO, 2013).

Nesta pesquisa, trabalhamos com classificação textual (*text classification*), que é um tipo de aprendizagem supervisionada, pois há intervenção humana na definição prévia das categorias. Em nosso cenário de aplicação, utilizamos as categorias pré-definidas dos GTs da Ancib. Nosso objetivo de classificar documentos de textos (resumos e textos completos) das teses e dissertações, de acordo com categorias pré-definidas, pode descrever a tarefa como uma função: $DxC\ \{T,F\}$, onde $D=\{d1,d2,...,d223\}$ é o conjunto que representa o *corpus* de documentos, no nosso caso 223 documentos de teses e dissertações, e $C=\{c1,c2,...,c11\}$ é o conjunto pré-definido de categorias que são os 11 GTs da Ancib.

O valor *T* atribuído a <*dj,ci*> indica uma decisão de classificar *dj* como *ci*, e *F* indica que *dj* não é classificado como *ci* (BAEZA-YATES; RIBEIRO-NETO, 2013). A figura 1 esquematiza os módulos que constituem a técnica proposta para classificação automática de teses e dissertações.

[7] Zip (ou ZIP) é um formato de arquivo usado para compactação de dados armazenados no computador. O objetivo da compactação é reduzir o tamanho de um arquivo ou agrupar vários arquivos em um só.

Figura 1 − Proposta para a classificação automática de teses e dissertações da CI

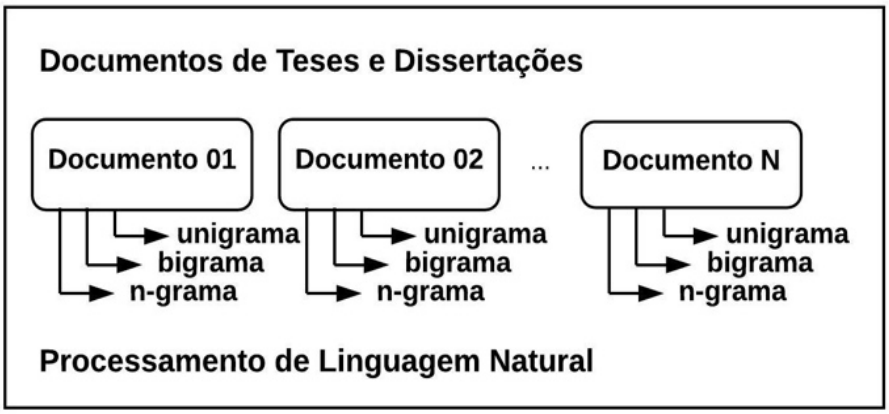


Fonte: Dyck, Dutra e Viera (2019).

1. **Repositório Institucional:** Ambiente de acesso livre e irrestrito à literatura científica e acadêmica da instituição.
2. **Documentos de Teses e Dissertações:** Os documentos (resumos e textos completos) extraídos do Repositório Institucional submetidos ao processamento de linguagem natural.
3. **Classificador:** Onde ocorre o treinamento, o cálculo do grau de pertencimento e a obtenção da função de classificação.
4. **Categorias / Ementas dos Grupos de Trabalho da Ancib:** As ementas dos GTs da Ancib extraídos do site do "Fórum de Coordenadores de Grupo de Trabalho da Ancib", submetidos a processamento de linguagem natural.
5. **Fórum de Coordenadores de Grupo de Trabalho da Ancib:** Site com a apresentação da Ancib e a descrição e ementas de cada um dos seus GTs.
6. **Documentos classificados em Grupo de Trabalho da Ancib:** Resultados da classificação automática com os documentos classificados em uma categoria (um GT da Ancib).

A partir da interação destes seis módulos, chega-se à classificação automática dos documentos das teses e dissertações com relação aos GTs da Ancib. O módulo 1 representa o RI, ambiente de acesso livre e irrestrito à literatura científica e acadêmica da instituição, que é a fonte dos documentos das publicações de teses e dissertações.

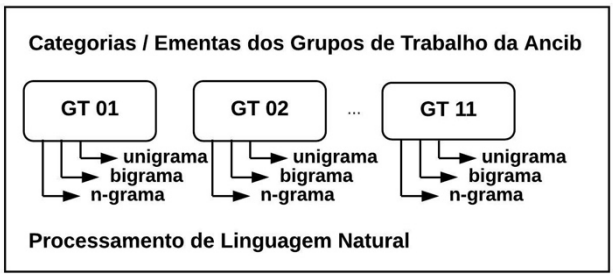
Figura 2 − Processamento de linguagem natural sobre os documentos de teses e dissertações



Fonte: Dyck, Dutra e Viera (2019).

O módulo 2, detalhado na figura 2, apresenta o processamento de linguagem natural, utilizando o modelo BoW, para a criação e limpeza de dados dos n-gramas, para cada um dos documentos, resumos e textos completos, de teses e dissertações extraídos do RI. O módulo 3 representa a classificação automática, no qual ocorre o treinamento, o cálculo do grau de pertencimento e a obtenção da função de classificação.

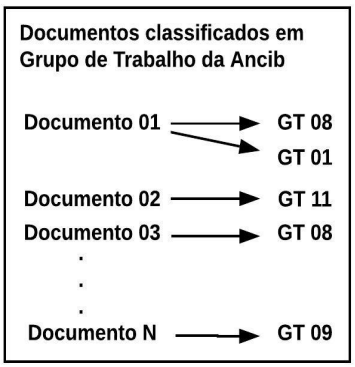
Figura 3 − Processamento de linguagem natural sobre as Categorias / Ementas dos GTs da Ancib



Fonte: Dyck, Dutra e Viera (2019).

O módulo 4, detalhando na figura 3, apresenta o processamento de linguagem natural, semelhante ao módulo 2, utilizando o modelo BoW, para a criação e limpeza de dados dos n-gramas, para cada uma das categorias / ementas dos 11 GTs da Ancib. O módulo 5 representa o "Fórum de Coordenadores de Grupo de Trabalho da Ancib", do qual foram extraídas as categorias / ementas de cada um dos GTs. O módulo 6, detalhado na figura 4, apresenta o resultado da classificação automática, em que cada um dos documentos de teses e dissertações foi classificado com uma probabilidade de pertencimento a determinado GT da Ancib.

Figura 4 − Documentos classificados em GTs da Ancib



Fonte: Dyck, Dutra e Viera (2019).

Nosso cenário de aplicação foi montado sobre as onze ementas dos GTs da ANCIB e os 223 documentos das teses e dissertações em CI, do PGCIN, obtidos no RI/UFSC. Uma vez de posse dos 223 documentos, os seus metadados nos mostraram que, de fato, tínhamos 185 resumos. E, de posse do arquivo ZIP com todos os textos completos, tínhamos 190 textos completos, aptos a serem trabalhados. Os documentos restantes não possuíam resumo e/ou texto completo hospedado no RI/UFSC. Nosso processo, desenvolvido para identificar automaticamente o grau de pertencimento dos documentos, conta com as seguintes etapas:

1. Criamos uma lista de *stop words* (palavras sem valor semântico), a partir das *stop words* relacionadas dentro da lista do NLTK[8], que é uma biblioteca do Python;

2. Geramos os dicionários com as ementas dos GTs. Para cada ementa, pegamos todo o seu texto e o colocamos em uma variável de dicionário. Pegamos, então, esse texto da ementa e o colocamos dentro de variáveis do tipo STRING[9]. Isto é, pegamos todo o texto original da ementa sem qualquer alteração. É esse texto que está na variável. Assim, temos uma variável para cada GT;

3. Aplicamos uma função, que criamos, que é aplicada sobre todas essas variáveis, para remover toda acentuação;

4. Em seguida, passamos para o processo de normalização em cada um dos documentos (ementas, resumos e textos completos). Retiramos a acentuação, caracteres que não são de A-Z, com ou sem acento. Transformamos todos os caracteres acentuados em não acentuados. Isso também é feito para os caracteres latinos, como a letra "Ç", que é transformada na letra "C".

[8] https://www.nltk.org/

[9] O termo STRING serve para identificar uma sequência de zero ou mais caracteres. Na prática, as STRINGS são usadas para representar textos.

Retiramos, então, todas as pontuações-padrão ('.', ',', ';', '!', '?', etc.), e transformamos todas as letras em maiúsculas. Esse mesmo processo de normalização também é aplicado à lista de *stop words*, criada inicialmente, a partir da lista do NLTK. Até aqui, temos nossas variáveis do tipo STRINGS normalizadas;

5. Então, as variáveis STRING são divididas (*splitted*) e são geradas listas, que são os vetores de palavras. A partir deste momento, o que é feito tem relação com o processo de mineração de dados[10]. As variáveis são colocadas em dicionários, que são uma estrutura de dados que possui os campos 'chave' e 'valor'. O campo chave armazena o número do GT e o campo valor armazena um vetor de palavras, para cada GT, criando-se um *set* com eles. Criar um *set* consiste em retirar as palavras repetidas. O *set* garante que as palavras restantes são únicas. É preciso garantir que estas palavras sejam únicas, que não se repitam, para não se gerar métricas equivocadas. Elas serão utilizadas na comparação com os documentos de resumos e textos completos de teses e dissertações. Essa lista de vetores de palavras está contida dentro do dicionário dos GTs;

6. A seguir, carregamos os resumos, numa primeira rodada, e os textos completos, numa segunda rodada. Neste momento, aplicamos uma função, chamada "VerificaTaxa", apresentada na figura 5, para identificar o grau de pertencimento entre um documento (resumo ou texto completo) um GT da Ancib. Essa verificação se repete para todos os textos, que, neste cenário, são as ementas dos GTs, os resumos e os textos completos dos documentos de teses e dissertações. Isso é necessário para podermos compará-los. A função verifica, para todas as palavras de uma ementa de GT, sua existência, ou não, na lista de palavras dos documentos (resumo ou texto completo). A seguinte verificação é feita:

a) Quantas dessas palavras existem dentro desse documento? Quantas vezes aparece essa palavra dentro do documento? Nenhuma, duas, três, etc., e incrementa-se um contador para o GT em avaliação. É utilizado um contador para cada GT. É contabilizada a quantidade de palavras do GT, encontradas dentro de um documento (resumo ou texto completo);

b) Depois de verificar todo o documento e chegar em um número final de ocorrências de palavras do vetor do GT, aplicamos a fórmula abaixo, na qual, pega-se esse número de ocorrências e divide-se pelo número total de palavras (tamanho total, *length*) do documento (resumo ou texto completo) e multiplica-se por 100, para verificar a porcentagem, ou seja, o grau de pertencimento do GT dentro daquele documento. Caso o resultado seja ZERO, automaticamente, sabemos que não há pertencimento do documento ao GT com o qual foi feita a comparação. O resultado é diferente de ZERO quando ao menos uma mesma palavra foi encontrada tanto no GT como no documento. E, nesse caso, registramos numa variável chamada "tupla" os valores: número do GT e grau de pertencimento, conforme figura 5.

$$\frac{Número\ de\ ocorrências}{Número\ total\ de\ palavras} \times 100 = Grau\ de\ pertencimento$$

Figura 5 – Código Python da função "VerificaTaxa"

```
def VerificaTaxa(texto):
    texto = remover_acentos(texto)
    texto = texto.split()
    texto = RemoveStopWord(texto)

    resultado = list()
    for gt in grupos:
        lista_frequencia = list()
        quantidade = 0
        for palavra in grupos[gt]:
            quantidade = quantidade + texto.count(palavra)

        if quantidade>0:
            freq_gt = (quantidade / len(texto))*100
            tupla = (gt, freq_gt)
            resultado.append(tupla)

    return resultado
```

Fonte: Autores (2020).

[10] Mineração de dados (também conhecida pelo termo inglês data mining) é o processo automatizado ou semiautomatizado para extrair conhecimentos e padrões, a partir de grandes quantidades de dados (OLAFSSON; LI; WU, 2008).

O processo de comparação entre os n-Gramas das ementas dos GTs com os resumos, num primeiro momento, e com os textos completos, dos documentos de teses e dissertações, num segundo momento, diversifica a quantidade de palavras a serem comparadas, de acordo com a variação do número **n** do **n**-grama utilizado, da seguinte maneira:

- Para o n = 1, os unigramas, os vetores são constituídos pelas próprias palavras, sem a necessidade de se aplicar a função para a formação dos n-gramas maiores que 1, conforme a seguir;
- Para o n = 2, os bigramas, a função inicia pegando a primeira e segunda palavras do documento. Assim, o primeiro bigrama é composto pela primeira e segunda palavras do documento. Em seguida, a função pega a segunda palavra do documento, junto com a palavra seguinte, que, neste caso, é a terceira. Agora, este novo bigrama é composto pela segunda e terceira palavras do documento. E, assim, sucessivamente, vai para a próxima palavra e pega também a palavra seguinte, e repete este processo até o final do documento;
- Para o n = 3, os trigramas, a função inicia pegando a primeira, a segunda e a terceira palavras do documento. Assim, o primeiro trigrama é composto pela primeira, segunda e terceira palavras do documento. Em seguida, a função pega a segunda palavra do documento, junto com as duas palavras seguintes, que são, agora, a terceira e a quarta palavras do documento. E, assim, sucessivamente, vai para a próxima palavra e pega também as duas palavras seguintes e repete este processo até o final do documento;
- Para o n = 4, a função inicia pegando a primeira, segunda, terceira quarta palavras do documento. Assim, o primeiro n-grama = 4 é composto pela primeira, segunda, terceira e quarta palavras do documento. Em seguida, a função pega a segunda palavra do documento, junto com as três palavras seguintes, que são, agora, a terceira, a quarta e a quinta palavras do documento. E, assim, sucessivamente, vai para a próxima palavra e pega também as três palavras seguintes, e repete este processo até o final do documento;
- E, para o n=5, a função inicia pegando a primeira, a segunda, a terceira, a quarta e a quinta palavras do documento. Assim, o primeiro n-grama = 5 é composto pela primeira, segunda, terceira, quarta e quinta palavras do documento. Em seguida, a função pega a segunda palavra do documento, junto com as quatro palavras seguintes, que são, agora, a terceira, a quarta, a quinta e a sexta palavras do documento. E, assim, sucessivamente, vai para a próxima palavra e pega também as quatro palavras seguintes, e repete este processo até o final do documento.

Isso é feito tanto para os vetores com as palavras das ementas dos GTs, como para os vetores com as palavras dos documentos, sejam os resumos ou os textos completos. É feito para todos os documentos: ementas, resumos e textos completos.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

O resultado da aplicação do processo, no cenário de aplicação aqui apresentado é sintetizado nas tabelas 1 e 2. A primeira com os resultados da aplicação do processo sobre os resumos e a outra com os resultados da aplicação do processo sobre os textos completos. Considerando que o número total de resumos avaliados foi de 185, observamos que os resultados da aplicação do processo, utilizando unigramas, n-grama = 1, sem *stop words*, que aparece na primeira linha de resultados da tabela 1, mostra um grau de pertencimento bem próximo do total de 185 resumos, em todos os GTs. No GTs 2, 6 e 7, o grau de pertencimento atingiu a totalidade dos 185 resumos. Na aplicação do processo, utilizando unigramas, vetores de palavras com n-grama = 1, ou seja, com uma única palavra a ser comparada, utilizamos apenas vetores sem *stop words*. Os resultados da aplicação do processo utilizando os bigramas, n-grama = 2, foram idênticos tanto para os vetores com e sem *stop words*. Esses resultados, da aplicação de unigramas e bigramas na comparação dos termos, obtendo valores extremos, próximos da totalidade de documentos avaliados, neste cenário de aplicação, indicam não ser apropriado seu uso.

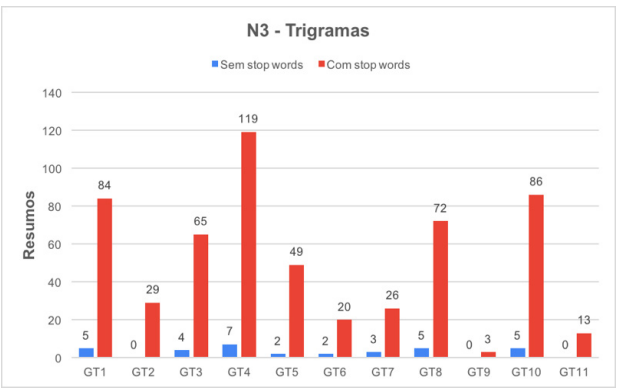
Tabela 1 – Resultados para os Resumos

Resultados para os Resumos												
N-Gramas		GT1	GT2	GT3	GT4	GT5	GT6	GT7	GT8	GT9	GT10	GT11
N-Grama = 1	Sem stop Words	184	185	182	181	177	185	185	178	179	180	178
N-Grama = 2	Sem stop words	170	167	140	180	150	165	151	158	148	162	154
	Com stop words	170	167	140	180	150	165	151	158	148	162	154
N-Grama = 3	Sem stop words	5	0	4	7	2	2	3	5	0	5	0
	Com stop words	84	29	65	119	49	20	26	72	3	86	13
N-Grama = 4	Sem stop words	0	0	0	1	0	0	0	1	0	0	0
	Com stop words	35	7	28	41	6	1	5	17	1	37	1
N-Grama = 5	Sem stop words	0	0	0	0	0	0	0	0	0	0	0
	Com stop words	3	3	1	12	3	0	0	8	0	3	0

Fonte: Autores (2020).

A aplicação do vetor com trigramas, n-grama = 3, mostra resultados com uma significativa oscilação, não ficando em zero, quando sem o uso de *stop words*, exceto para os GTs 2, 9 e 11. Há, também, uma oscilação longe do extremo de todos os resumos, 185, com resultados variando entre grau de pertencimento igual a 3 para o GT9 e grau de pertencimento igual a 119 para o GT4, conforme figura 6. Esses resultados apresentam graus bem mais reduzidos quando se utilizam vetores sem *stop words*.

Figura 6 – Resumos em Trigramas: Número de documentos pertencentes a cada GT



Fonte: Elaborado pelos autores (2020).

A aplicação do vetor com n-grama = 4, sem o uso de *stop words,* obteve resultados de grau de pertencimento igual a zero, nas comparações com todos os GTs, exceto para os GTs 4 e 8, nos quais o resultado foi igual a um. Esses resultados também indicam não ser apropriado o uso de n-gramas sem *stop words* a partir de n = 4, uma vez que no cenário de aplicação aqui utilizado não identificam pertencimento a, praticamente, nenhum dos GTs avaliados. Isso se justifica pelo fato de que, quando se trata de n-gramas, as *stop words* adquirem a importância que não possuem na mineração textual com termos simples, devido ao fato de que, combinadas com outras palavras da *n-upla*, elas, neste caso, possuem valor semântico. A aplicação do vetor com n-grama = 4 com o uso de *stop words* obteve resultados que pareceram promissores, num primeiro momento, porém, quando comparados com a totalidade de resultados, com os graus de pertencimento em relação a todos os GTs, apresenta uma grande oscilação nos resultados, variando desde um grau de pertencimento igual a um, com o GTs 6, 9 e 11, até uma taxa de quarenta e um, com o GT4.

Já os resultados para os resumos, com n=5, tanto sem como com *stop words*, apresentam combinações (*matches*) consideravelmente reduzidas.

A tabela 2 mostra os resultados da aplicação do processo com os textos completos, no total de 190 documentos, de teses e dissertações do PGCIN/UFSC. Os resultados da aplicação do processo, utilizando unigramas sem *stop words*, e bigramas tanto com, quanto sem *stop words*, que aparecem nas três primeiras linhas da tabela 2, à semelhança dos resultados obtidos com os resumos, obtendo resultados da totalidade de documentos avaliados, neste cenário de aplicação, indicam não ser apropriado seu uso. Isto ocorre porque nos textos completos a probabilidade de ocorrência de termos simples (n = 1) ou em duplas (n = 2), ainda que desconectados da questão semântica, é muito maior, ou seja, estas duas situações acabam por não servir de balizamento para a determinação do grau de pertencimento.

Tabela 2 − Resultados para os Textos completos

Resultados para os Textos completos												
N-Gramas		GT1	GT2	GT3	GT4	GT5	GT6	GT7	GT8	GT9	GT10	GT11
N-Grama = 1	Sem stop Words	190	190	190	190	190	190	190	190	190	190	190
N-Grama = 2	Sem stop words	190	190	190	190	190	190	190	190	190	190	190
	Com stop words	190	190	190	190	190	190	190	190	190	190	190
N-Grama = 3	Sem stop words	48	47	48	47	46	42	44	47	34	47	42
	Com stop words	189	184	188	189	186	166	182	190	132	190	168
N-Grama = 4	Sem stop words	9	4	10	10	0	3	3	5	0	8	0
	Com stop words	182	78	165	182	83	48	127	167	16	177	15
N-Grama = 5	Sem stop words	3	0	0	4	0	0	1	1	0	1	0
	Com stop words	79	13	4	151	41	1	34	113	1	74	1

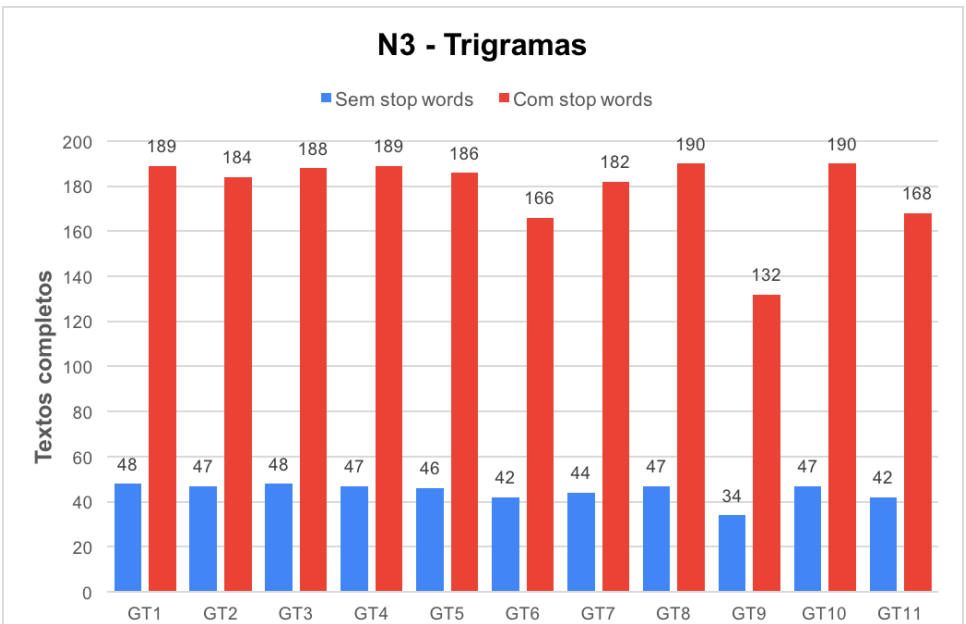
Fonte: Elaborado pelos autores (2020).

Com os resultados da aplicação do processo no vetor com trigramas, n-grama = 3, apresentado na figura 7, entramos numa possível discussão do "equilíbrio perfeito". Percebemos, com estes resultados, que existe um equilíbrio nos valores da aplicação sem *stop words*. Aqui, fazendo uma média aritmética dos resultados de cada um dos GT, sem *stop words*, e dividindo pelo total de documentos avaliados, 190, e multiplicando por 100, chega-se ao resultado de que 23% dos documentos fazem parte de algum GT. Isto é, 23% dos textos completos estão inseridos em cada GT. Este resultado mostra um equilíbrio na distribuição das teses e dissertações do PGCIN, nos 11 GTs da Ancib. Quase 1/4 dos documentos estão representados em cada GT.

Nos resultados da aplicação de trigramas com *stop words*, os valores obtidos para os GTs 8 e 10 mostram o resultado de 190, que é todo universo de documentos analisados. Esses valores aparecem na linha 5 dos resultados apresentados na tabela 2. Para isso ter acontecido, é indicação de que trigramas com palavras sem valor semântico, como, por exemplo, "E ISSO E", "POR QUE ISSO", tenham sido utilizados na comparação. Uma vez que, em nosso processo, basta o contador ser igual a um, ou seja, existir apenas uma combinação (um *match*), para considerar que existe um pertencimento daquele documento ao GT. Ou seja, quando se mantém as *stop words*, temos, potencialmente, esse tipo de resultado. Quando retiramos as *stop words*, em nosso processo de comparações, significando a inexistência de palavras com baixo, ou nenhum, peso semântico, automaticamente, percebemos resultados com números menores nos graus de pertencimento, sugerindo maior confiabilidade.

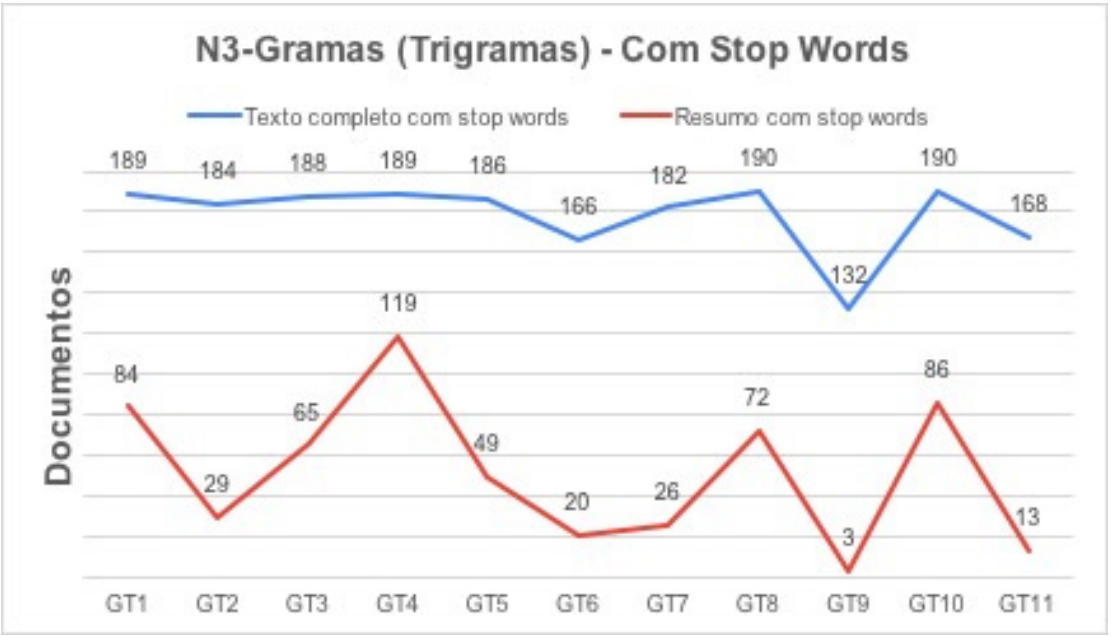
Os resultados, apresentados na tabela 2, mostram que a confiabilidade pode ser auferida com os trigramas sem *stop words*. Diante dos resultados obtidos nesta pesquisa, é a partir do n = 3, dos trigramas, que se obtém os melhores graus de pertencimento, com as melhores chances de acerto. A figura 8 mostra os resultados obtidos na aplicação de trigramas com *stop words*, tanto para resumos quanto para textos completos. A figura 9 mostra os resultados obtidos na aplicação, também de trigramas sem *stop words* tanto para os resumos e textos completos.

Figura 7 – Textos completos em Trigramas: número de documentos pertencentes a cada GT



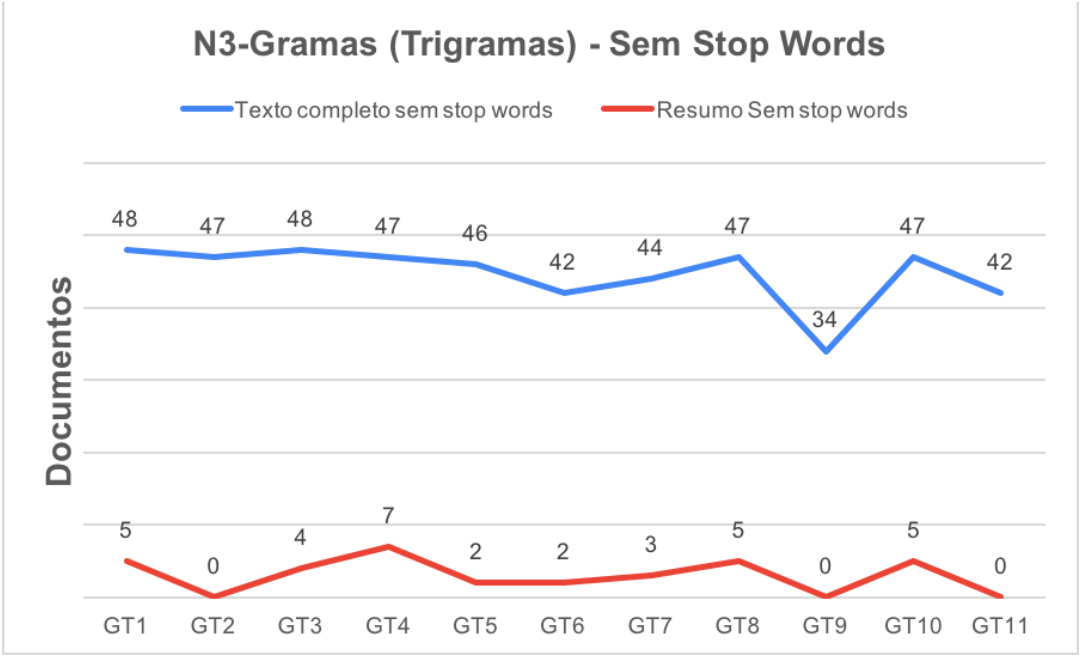
Fonte: Elaborado pelos autores (2020).

Figura 8 – Resultados dos Trigramas com stop words para Resumos e Textos completos



Fonte: Elaborado pelos autores (2020).

Figura 9 – Resultados dos Trigramas sem stop words para Resumos e Textos completos



Fonte: Elaborado pelos autores (2020).

CONCLUSÕES

Esta pesquisa utilizou um tipo de aprendizagem supervisionada para classificar textos de maneira geral, de acordo com categorias pré-definidas. Utilizamos como cenário de aplicação o universo de documentos de teses e dissertações do PGCIN/UFSC, de acordo com as ementas dos GTs da ANCIB. Utilizando o modelo BoW, propusemos uma arquitetura de classificação automática, identificando o grau de pertencimento de cada um dos documentos. Os resultados das comparações com os textos completos, a partir da utilização de n-grama = 3, demonstraram que a utilização de classificação por n-gramas implica um cenário em que quanto maior o número de gramas, sem *stop words*, menor será grau de pertencimento obtido. Ou seja, será muito mais difícil se identificar uma combinação entre documentos (*match*). É preciso, no entanto, que seja levada em consideração a semântica do cenário de aplicação trabalhado na decisão de se incluir ou não as *stop words* na formação dos n-gramas e serem buscados. Dependendo do contexto, n-gramas conhecidos e que representam expressões corriqueiras poderão se basear fortemente na utilização de elementos textuais, não obstante, considerados *stop words*. Será preciso achar um equilíbrio nas decisões de projeto de maneira a se maximizar o grau de pertencimento obtido, com a extração da maior semântica possível do corpus *textual*.

A proposta desta pesquisa foi de identificar o grau de pertencimento entre textos de maneira geral, utilizando como cenário de aplicação os documentos de teses e dissertações do PGCIN em relação aos GTs da ANCIB. Como premissa de pesquisa, é possível que um documento pertença, em maior ou menor grau, a vários GTs, pois, mesmo que a linha de pesquisa na qual a tese ou dissertação foi escrita, focada num assunto particular, que tenha relação com a ementa de um determinado GT, considerando que a área da Ciência da Informação possui inúmeras interseções com diferentes áreas do conhecimento, sendo multidisciplinar, é possível referenciar, num único trabalho, assuntos que são fortemente trabalhados em diferentes GTs. Isso também pode acontecer devido ao que é discutido na introdução dos trabalhos, muitas vezes fazendo apresentações históricas da área. Isso sugere, portanto, a necessidade de um trabalho em que seja definida uma linha de corte, para se definir o grau de pertencimento a um GT. A definição de uma linha de corte apropriada demandará um trabalho criterioso.

Outra questão a ser ponderada, foi a utilização das ementas dos GTs, encontradas no site da ANCIB, como único documento de fonte textual para caracterizar cada GT. Para termos maior confiança nos resultados, é necessário um *corpus* maior, para ampliar o número de palavras que caracterizem cada GT. Que então, por sua vez, serão utilizadas para a classificação das teses e dissertações.

Futuros trabalhos incluem a definição de um limite de corte, no número de palavras coincidentes, para então considerar que existe um pertencimento a determinado GT; comparar os resultados da classificação dos resumos e seus respectivos textos completos, de um mesmo documento, tese ou dissertação no cenário aqui utilizado, e comparar os resultados do grau de pertencimento, para então avaliar a representatividade dos resumos em relação aos seus textos completos; o desenvolvimento de um aplicativo, de livre acesso pela internet, que permita classificação de quaisquer *corpora* de documentos de acordo com categorias pré-definidas; testar o modelo com um método não-supervisionado, a Clusterização, usando volumes maiores de dados. E, também, posteriormente, substituir o método BoW por uma técnica que preserve a semântica, como, por exemplo, os vetores de palavras (WordEmbeddings).

REFERÊNCIAS

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Recuperação de Informação: conceitos e tecnologia das máquinas de busca.* 2. Ed. Porto Alegre: Bookman, 2013.

DYCK, A. F.; DUTRA, M. L.; VIERA, A. F. G. Classificação automática de teses e dissertações da área da Ciência da Informação sob a ótica dos grupos de trabalho da Ancib. In: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 2019, Brasília. *Anais* [...] Widat 2019. Brasília: Editora da UnB, 2019. p. 48-53.

FAN, W. *et. Al*. Tapping the power of text mining. *ACM,* [*s. l.*], v. 49, n. 9, p. 76-82, 2006. DOI: https://doi.org/10.1145/1151030.1151032

GOLDBERG, Y. *Neural network methods in natural language processing*: synthesis lectures on human language technologies. [*S. l.*]: Morgan & Claypool Publishers. 2017. 310 p.

HARRIS, Z.S. Distributional Structure. *WORD*, [*s. l.*], v. 10, n. 2-3, p. 146-162, 1954. Publicado online em 04 dez. 2015. ISSN: 0043-7956. Disponível em: <https://www.tandfonline.com/doi/pdf/10.1080/00437956.1954.11659520> Acesso em: 11 fev. 2020.

INGERSOLL, G.S.; MORTON, T.S.; FARRIS, A.L. 2013. *Taming text:* how to find, organize and manipulate it. Shelter Island, NY (USA): Manning Publications Co., 2012. 298 p. ISBN: 9781933988382

JURAFSKY, D.; MARTIN, J.H. S*peech and language processing*: an introduction to natural language processing, computational linguistics, and speech recognition. 3. ed. Stanford University. 2020. Disponível em:<https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>. Acesso em: 04 mar. 2020.

KUMAR, P. An introduction to n-grams: what are they and why do we need them?. *XRDS Crossroads The ACM Magazine for Students.* 2017. Disponível em: <https://blog.xrds.acm.org/2017/10/introduction-n-grams-need/>. Acesso em: 11 fev. 2020.

MEIRELES, M. R. G.; CENDóN, B. V. Categorização e classificação de documentos a partir de suas citações: uma proposta baseada em redes neurais artificiais. *DataGramaZero*, v. 12, n. 5, 2011. Disponível em: <http://hdl.handle.net/20.500.11959/brapci/7466>. Acesso em: 11 mar. 2021.

MOURA, M.F.; *et al. Um modelo para seleção de n-gramas significativos e não redundantes em tarefas de mineração de textos.* Campinas: Embrapa Informática Agropecuária, 2010. (Boletim de pesquisa e desenvolvimento, n. 23). ISSN 1677- 9274.

OLAFSSON, S.; LI, X.; WU, S. Operations research and data mining. *European Journal of Operational Research*, [*s. l.*], v. 187, n. 3, p. 1429-1448. 2008. ISSN 0377-2217. DOI: https://doi.org/10.1016/j.ejor.2006.09.023.

SILVA, E.L. *Metodologia da pesquisa e elaboração de dissertação.* Estera Muszkat Menezes. 4. ed. rev. atual. Florianópolis: UFSC, 2005. 138p.

Dados e metadados: conceitos e relações

Ana Carolina Simionato Arakaki
Doutora em Ciência da Informação pela Universidade Estadual Paulista (Unesp) - Marília, SP - Brasil.
Professor da Universidade Federal de São Carlos (UFSCar) - São Carlos, SP - Brasil.
http://lattes.cnpq.br/9896600626524397
E-mail: acsimionato@ufscar.br

Felipe Augusto Arakaki
Doutor em Ciência da Informação pela Universidade Estadual Paulista (Unesp) - Marília, SP - Brasil.
Professor da Universidade de Brasília (UnB) - Brasília, DF - Brasil.
http://lattes.cnpq.br/5324289839207169
E-mail: fe.arakaki@gmail.com

Submetido em: 25/11/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021.

RESUMO

A interdisciplinaridade das pesquisas em áreas correlatas pode apresentar muitas contribuições para o desenvolvimento de teorias. Entretanto, também pode proporcionar confusões terminológicas. Dessa forma, destaca-se a importância em discutir esses conceitos e evitar equívocos de nomenclatura. Por essa razão, o objetivo deste trabalho consiste em discutir e relacionar conceitos de dados e metadados na Ciência da Informação. A pesquisa é caracterizada por uma metodologia de análise exploratória, sendo possível identificar elementos conceituais por meio da literatura científica, a analisar as informações a partir do Perspectivismo. Como resultados, são apresentadas as definições e as relações entre 'dado' e 'metadado' identificadas na revisão da literatura. Considera-se que o levantamento e as discussões apontadas no texto demonstram diversas relações entre os conceitos, principalmente, atrelados à *Web* Semântica, Ciência dos dados, *Big data*, entre outros, o que leva à necessidade de aprofundamento nas definições e reflexões na área de Ciência da Informação, devido às divergências de perspectivas conceituais.

Palavras-chave: Dados. Metadados. Ciência da Informação. Dado e metadado.

Data and metadata: concepts and relationships

ABSTRACT

The interdisciplinarity of research in related areas can present many contributions to the development of theories. However, it can also provide terminological confusion. Thus, the importance of discussing these concepts and avoiding nomenclature misconceptions is highlighted. For this reason, the objective of this work is to discuss and relate data and metadata concepts in Information Science. The research is characterized by an exploratory analysis methodology, making it possible to identify conceptual elements from the scientific literature, analyzing the information from Perspectivism. As a result, the definitions and relationships between 'data' and 'metadata' identified in the literature review are presented. It is considered that the survey and the discussions presented in the text demonstrate several relationships between the concepts, mainly linked to the Semantic Web, Data Science, Big data, among others, raising the need for further definitions and reflections in the area of Science of the Information, due to divergences in conceptual perspectives.

***Keywords**: Data. Metadata. Information Science. Data and metadata.*

Datos y metadatos: conceptos y relaciones

RESUMEN

La interdisciplinariedad de la investigación en áreas relacionadas puede presentar muchas contribuciones al desarrollo de teorías. Sin embargo, también puede provocar confusiones terminológicas. Por lo tanto, se destaca la importancia de discutir estos conceptos y evitar conceptos erróneos de nomenclatura. Por esta razón, el objetivo de este trabajo es discutir y relacionar los conceptos de datos y metadatos en Ciencias de la Información. La investigación se caracteriza por una metodología de análisis exploratorio, que permite identificar elementos conceptuales de la literatura científica, analizando la información del Perspectivismo. Como resultado, se presentan las definiciones y relaciones entre "datos" y "metadatos" identificados en la revisión de la literatura. Se considera que la encuesta y las discusiones presentadas en el texto demuestran varias relaciones entre los conceptos, principalmente vinculados a la Web Semántica, la Ciencia de Datos, Big Data, entre otros, lo que plantea la necesidad de nuevas definiciones y reflexiones en el área de la Ciencia de la Información, debido a divergencias en perspectivas conceptuales.

***Palabras clave**: Datos. Metadatos. Ciencia de la información. Datos y metadatos.*

INTRODUÇÃO

Diversas áreas apresentam contribuições significativas para a construção e expansão da *Web*. Essa pluralidade de teorias e aplicações para a *Web* gera confusões terminológicas entre áreas correlatas cada vez mais frequentes. Por essa razão, é necessário refletir sobre os conceitos, no intuito de amenizar ou evitar as ambiguidades com relação ao vocabulário conceitual.

O termo dado, por exemplo, foi empregado em diversos contextos, principalmente na perspectiva do atual cenário e das discussões envoltas nas questões da *Web* semântica, *Linked Data*, *Big Data*, curadoria digital, *e-Science*. Como também, o termo foi relacionado entre outros termos e conceitos nas tendências da Ciência de Dados (*Data Science*), Ciência da Computação e Ciência da Informação.

Nesse sentido, o uso equivocado pode causar confusões em teorias, hipóteses, metodologias e no próprio desenvolvimento científico. Assim, destaca-se a importância de se contextualizar os conceitos e objetos nos estudos, quando é trabalhado com termos que possuem diversas definições e concepções entre duas ou mais áreas.

Na área de Ciência da Informação, isso não é diferente. Com forte relacionamento e interdisciplinaridade com a Ciência da Computação, Ciência de Dados, dentre outras áreas, o termo dado tem tomado inúmeros significados. No caso, foi observado pela literatura que o termo tem sido empregado como sinônimos de metadados, e ainda, na atualização e evolução dos conceitos relacionados a esses dois termos (dados e metadados).

Furner (2019) aponta a necessidade de discutir e comparar o conceito de metadados com outros termos, como exemplo: dados, documento, informação e os dados do próprio registro. Além disso, Baker *et al.* (2011) ressaltam a necessidade de uma discussão conceitual da terminologia empregada em áreas interdisciplinares, como a Ciência da Informação e a Ciência da Computação.

Diante desse cenário, o objetivo deste trabalho consiste em discutir e relacionar os conceitos dos termos 'dado' e 'metadado', no intuito de debater a proximidade conceitual na Ciência da Informação. Portanto, com a apresentação desse texto, busca-se compartilhar as reflexões do Grupo de Pesquisa "Dados e Metadados" sobre a relevância em contextualizar e discutir sobre esses conceitos na literatura.

Apresenta-se o artigo em uma estrutura dividida em: introdução, uma seção para os procedimentos metodológicos adotados, a análise e discussão sobre dados e metadados divididos em duas seções e, por fim, as considerações finais.

PROCEDIMENTOS METODOLÓGICOS

A pesquisa é caracterizada por uma análise exploratória para identificar os elementos conceituais que visam, a partir da literatura científica da área de Ciência da Informação, os conceitos de dados e metadados.

A análise exploratória foi realizada pelo estudo sobre dados e metadados por meio de uma pesquisa bibliográfica. O recorte da pesquisa abrange outras já publicadas internacionalmente e no Brasil, nos idiomas: português, espanhol e inglês. Os procedimentos metodológicos foram divididos em quatro etapas, conforme descrito a seguir:

A primeira etapa está relacionada ao levantamento bibliográfico, que foi realizado em bases de dados como: Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI); Repositório Questões em Rede – Coleções (Coleção BENANCIB); *Library and Information Science Abstracts* (LISA); *Library, Information Science and Technology Abstracts* (LISTA); Biblioteca Digital Brasileira de Teses e Dissertações (BDTD); Teses e dissertações da CAPES; OASISBR: portal brasileiro de publicações científicas em acesso aberto do Ibict; Portal de Periódicos da Capes; *Scientific Electronic Library Online* (SciELO); *Scopus*; e *Web of Science*. Além de outras fontes de informação, como os Anais do *Dublin Core Metadata Initiative International Conference on Dublin Core and Metadata Applications*; *Google Scholar* e as publicações do *World Wide Web Consortium* (W3C).

Para a realização do levantamento, elaborou-se a estratégia de busca com termos principais (dados e metadados) para a localização dos materiais selecionados. Quando possível, fez-se a busca pelo título, por palavras-chave e resumos, na última década, ou seja, 2008-2018.

A tipologia dos trabalhos pesquisados foi: artigos, livros, teses, dissertações, trabalhos em eventos e publicações da W3C.

Após a realização do levantamento e da identificação do *corpus* do trabalho, operou-se a segunda etapa, com uma leitura do resumo e, quando necessário, uma leitura prévia do texto para que pudesse aplicar os seguintes critérios para seleção do material para fundamentação teórica do texto: 1) a partir de uma leitura do resumo verificou-se a relevância da temática do artigo para o escopo da pesquisa; 2) idioma dos documentos (português, inglês e espanhol); e 3) atualidade e pertinência do texto para pesquisa.

A terceira etapa caracterizou-se pela leitura e pelo fichamento dos textos completos. Essa etapa teve como propósito o desenvolvimento da base teórica para a discussão dos diferentes pontos de vista identificados na literatura sobre o tema.

Logo em seguida, durante a quarta etapa, considerou-se o Perspectivismo proposto por Peterson (1996) para analisar os diversos pontos de vista dos conceitos identificados, pois,

> [...] a adequação de uma perspectiva é sempre relativa, mas ela não se estabelece de modo arbitrário. Na escolha de uma perspectiva deve-se considerar a sua adequação ao usuário a uma situação determinada, a um processo, a uma ontologia e a uma meta específica. É a análise de cada um destes componentes que se pretende desenvolver no estudo dos processos que atuam nas diretrizes, modelagens e estruturas de sistemas para atendimento de necessidades de sujeitos em ambientes informacionais específicos (SANTOS; VIDOTTI, 2009, *não paginado*).

Com essa perspectiva, na quarta etapa, caracterizou-se a análise e o estabelecimento das características fundamentais extraídas da literatura sobre dados e metadados, para elucidação do problema de pesquisa. Criou-se, assim, a base teórica para a elaboração (redação) da pesquisa.

DISCUSSÃO CONCEITUAL SOBRE OS METADADOS

Ao longo dos anos, autores como Alves (2010), Alves e Santos (2013), Joudrey, Taylor e Wisser (2018), Méndez Rodríguez (2002), Pomerantz (2015) e Zeng e Qin (2008, 2016) têm discutido e estruturado definições sobre o termo metadados na perspectiva da Ciência da Informação. Dentre esses autores, o que é possível identificar como ponto em comum é que o conceito de metadados está atribuído a uma informação estruturada para as ações de identificação, descoberta, seleção, uso, acesso e gerenciamento.

O termo metadado é conceituado por meio das "[...] informações de valor agregado que criam para organizar, descrever, rastrear e melhorar o acesso a objetos de informação e itens físicos e coleções, relacionados a esses objetos" (GILLILAND, 2016, p. 2, *tradução nossa*). Zeng e Qin (2016) indicam as variações de definição do conceito de metadados por meio de diferentes comunidades de prática, ou ainda, que metadados são informações sobre coisas específicas.

Os metadados, inicialmente, foram identificados pela expressão 'dados sobre dados', cunhada na década de 1960 para se referir a um conjunto de declarações sobre os dados (POMERANTZ, 2015). No entanto, é perceptível que os metadados fazem parte da rotina de diversas classes profissionais que projetam, criam, descrevem, preservam e usam sistemas e recursos informacionais (GILLILAND, 2016).

Ressalta-se que, por meio dessa rotina, os metadados são ora entendidos como produtos, ora como processo. Mayernik (2020, *não paginado, tradução nossa*) apresenta em seu trabalho um panorama sobre os metadados. Destaca que "[...] o uso de padrões para criar metadados estruturados resulta no que pode ser caracterizado como 'produtos de metadados'". Isto é, o registro informacional gerado é o produto de metadados.

Edwards *et al.* (2011, p. 684, *tradução nossa*) esclarecem ainda que "[...] os produtos de metadados bem codificados aumentam a precisão com a qual um conjunto de dados que pode ser ajustado para finalidades para as quais não foi originalmente planejado ou pode ser reutilizado por pessoas que não participaram da criação". Por outro lado, o processo, por exemplo, pode ser relacionado como a interpretação de regras de instrumentos de representação, como o *Anglo-American Cataloguing Rules 2nd ed.* (AACR2r) no domínio bibliográfico, para apoiar no processo de decisão do catalogador durante a definição de informações que serão inseridas para compor o registro informacional.

Nesse sentido, Haynes (2004) afirma que a complexidade de compreensão dos metadados e as suas funções são essenciais para as atividades dos setores que envolvem conhecimento, informação, cultura e aprendizagem.

As funções dos metadados estão direcionadas à sua tipologia. Segundo Michael Buckland (2017), o primeiro e original uso dos metadados é a descrição de documentos, enquanto o uso adicional dos metadados é permitir a pesquisa. "Os metadados podem ser usados para fornecer estruturas que suportam a pesquisa e a descoberta consistentes de informações em uma ampla variedade de documentos. Os metadados também podem, potencialmente, permitir distinções entre tipos ou documentos semelhantes" (BUCKLAND, 2017, p. 118, *tradução nossa*).

A partir da análise terminológica realizada por Arakaki (2019), observou-se que, durante os últimos anos, houve uma disparidade na definição das tipologias dos metadados. Mayernik (2020) discute que essa variedade de categorias encontrada na literatura indica uma compreensão ampla do que o conceito de metadados pode abranger, além de representar tarefas ou ações específicas.

Para Mayernik (2020), a única semelhança entre essas várias definições e categorizações de metadados é que estes são criados para serem usados, por alguns propósitos e, ainda, por pessoas ou pelos aplicativos de computador. De acordo com Arakaki (2019, p. 80-81), as tipologias são identificadas como:

- *Metadados administrativos:* usados para gerenciar e administrar coleções e recursos informacionais, para auxiliar na tomada de decisão e na manutenção dos registros e recursos informacionais. Fornecem informações sobre a origem e a manutenção de um objeto;
- *Metadados de autenticação:* são informações que possibilitam a identificação, integridade, legitimidade de um recurso informacional;
- *Metadados de preservação:* estão relacionados com informações de preservação e conservação dos recursos informacionais;
- *Metadados de proveniência:* estão relacionados às informações de procedência, fornecem dados sobre entidades, criação e modificações e seus relacionamentos;
- *Metadados técnicos:* estão relacionados a como um sistema funciona, fornecem informações do sistema ou do recurso;
- *Meta-metadados (Metametadata):* correspondem às informações sobre o registro criado, ou informações da criação de um conjunto de dados;
- *Metadados descritivos:* identificam características identificadoras e os contextos intelectuais dos recursos de informação para fins de descoberta, identificação, seleção, aquisição, contexto e compreensão;
- *Metadados de direitos:* estão relacionados às informações sobre propriedade e direitos autorais;
- *Metadados de acesso e uso:* são informações de como um recurso informacional foi acessado e utilizado, como restrições de circulação e acesso, registros de exposições, entre outros;
- *Metadados estruturais:* estão relacionados à composição e à organização do recurso informacional;
- *Linguagens de marcação (Markup languages):* integram metadados e sinalizações para outros recursos estruturais ou semânticos.

As tipologias dos metadados estão presentes na seleção do padrão de metadados a ser utilizado no sistema informacional. Conforme apontado por Zeng e Qin (2008), há uma intrínseca ligação no estabelecimento de metadados e formatos de metadados. A construção de um padrão de metadados exige a adoção de procedimentos metodológicos para a definição dos metadados, assim como eles precisam estar em uma estrutura de descrição padronizada. Santos, Simionato e Arakaki (2014) apontam que a definição dos metadados deve ser uma ação consensual para que o sistema contenha interoperabilidade de seus dados e, ainda, quando há dados ambíguos e que necessitam de um elevado detalhamento do recurso informacional.

Ademais, os metadados comprovam a autenticidade e o grau de completude do recurso, estabelecem o seu contexto, identificam suas relações estruturais com outros recursos, provêm diversos pontos de acesso para diferentes tipos de usuários e podem fornecer informações que são geralmente obtidas por meios tradicionais (GILLILAND, 2016).

RELAÇÕES TERMINOLÓGICAS DO TERMO DADO

Os dados são destacados, inicialmente, pelo conceito atribuído por Santos e Sant'Ana (2013, p. 205), os quais definem dado como

> "[...] uma unidade de conteúdo necessariamente relacionada a determinado contexto e composta pela tríade entidade, atributo e valor, de tal forma que, mesmo que não esteja explícito o detalhamento sobre contexto do conteúdo, ele deverá estar disponível de modo implícito no utilizador, permitindo, portanto, sua plena interpretação".

A partir da definição de Santos e Sant'Ana (2014), observa-se que o dado é composto pela tríade entidade, atributo e valor. Apesar de implícito, o dado sempre está atrelado a um contexto.

No intuito de traçar uma evolução do conceito do termo dado, Furner (2016) faz um levantamento da definição de dados ao longo dos séculos. Por uma perspectiva histórica, discute a mudança do sentido. De acordo com o autor, o conceito dos dados pode possuir diversas abordagens, tais como:

- *Abordagem extensional:* busca caracterizar coisas ou tipos de coisas que se enquadram no conceito de "dados";
- *Abordagem intencional:* identifica as propriedades que algo deve ter, caso seja tratado como dados;
- *Abordagem classificatória:* reconhece um conceito individual como "dados" pode ter, ou possuir múltiplos sentidos, e que esses sentidos podem ser categorizados de acordo com similaridades em função e contexto;
- *Abordagem histórica:* conduz análises lógicas e/ou computacionais das propriedades necessárias dos conceitos, permitir que os autores considerem o desenvolvimento culturalmente específico dos significados de termos, como dados ao longo do tempo.

A partir dessas abordagens (extensional, intencional, classificatória e histórica), Furner (2016) destaca as diversas interpretações que os dados podem assumir:

- *A interpretação clássica:* origem do termo do latim *dătŭm*, como o verbo 'dar'. Furner (2016) discute a origem do termo em latim *dătŭm*, por volta do ano 100a.C., utilizado muitas vezes como verbo (dar);
- *A interpretação documental:* dados como metadados. De acordo com Furner (2016), ainda por volta do ano de 100a.C., o termo 'dado' (*dătŭm*) começou a ser usado como substantivo, como informação sobre algo, ou seja, como metadado;
- *A interpretação eclesiástica:* dados como dons de Deus. Furner (2016) relata que por volta de 1614, o termo 'dados' começou a ser utilizado em sermões, com o significado de 'com a graça de Deus';
- *A interpretação geométrica:* dados como premissas geométricas. No contexto da geometria, o termo 'dado' passou a ser empregado por volta de 1645, para representar os valores dos lados e ângulos (FURNER, 2016);
- *A interpretação matemática:* dados como premissas matemáticas. Após a interpretação geométrica, por volta de 1704, o conceito de dados foi ampliado para qualquer aplicação matemática, independente da área de aplicação (FURNER, 2016);
- *A interpretação epistêmica:* dados como evidência. Por volta de 1648, o termo 'dado' foi incorporado em alguns dicionários atribuindo o conceito de dados como fatos;
- *A interpretação informacional:* dados como valores de atributo. Na segunda metade do século XIX, há uma mudança na interpretação dominante do conceito de dados, sendo uma das principais alterações que o termo não apenas se refere a informações numéricas;
- *A interpretação computacional:* dados como *bits*. O termo 'dado' na computação foi aplicado pela primeira vez em 1953, com a publicação do IBM701 *Electronic Data Processing Machine.* O termo, porém, só foi constar nos dicionários terminológicos da computação em 1980. A princípio, utilizou-se o termo 'dado' para definir o valor do atributo em um banco de dados. Posteriormente, encontrou-se também como sinônimo de *bits*;
- *A interpretação diafórica:* por volta dos anos 2000, o termo relaciona-se à sua pluralidade. Os dados são atribuídos como realidade objetiva, aparências subjetivas, observações, ideias, significados, ou mesmo expressões linguísticas de observações individuais.

Ao analisar do ponto de vista da interpretação informacional dos dados destacada por Furner (2016), observa-se forte relação com a definição e a aproximação do conceito de metadados.

Diante disso, Mayernik (2020) discute a possibilidade de categorizar os metadados como uma tipologia dos dados e conclui que "[...] se os dados são entidades usadas como evidência, os metadados são os processos e produtos que permitem que essas entidades sejam responsabilizadas como evidência" (MAYERNIK, 2019, p. 735, *tradução nossa*).

Corroborando com essa discussão focada em ambientes de bibliotecas, Baker *et al.* (2011) atribuem aos dados que são produzidos por meio de uma informação ou curadoria. São nomeados como 'dados de biblioteca'. Ainda, descrevem recursos ou ajudam a sua descoberta. Os dados de bibliotecas dividem-se em: *datasets* (conjuntos de dados), *metadata element set* (conjuntos de elementos) e *value vocabularies* (vocabulários de valor) (BAKER *et al.*, 2011).

Os *datasets* são coleções estruturadas de metadados para descrição de recursos, como livro. Isto é, um conjunto de registros de metadados. Os *datasets* equivalem aos registros de bibliotecas que consistem em declarações, elementos da entidade e seus valores. Os elementos são definidos a partir de padrões, como *Machine Readable Cataloging* (MARC21) ou *Dublin Core* (DC) e os valores por vocabulários de valores, como a *Library of Congress Subject Headings* (LCSH) (ISAAC *et al.*, 2011). Pode-se considerar exemplos, como a *British National Bibliography*, o catálogo da *Hungarian National Library*, o *CrossRef* e a *Europeana* (BAKER *et al.*, 2011).

> Furner (2016, p. 303) argumenta que todos os conjuntos de dados são documentos, afirmando que 'o conjunto de dados é uma espécie de documento'. Portanto, se os metadados são um tipo especial de dados, conforme observado acima, os metadados também existem como documentos, não como conceitos abstratos ou informações que existem sem forma material. Como tal, os metadados podem ser analisados através do mesmo aparato conceitual dos documentos (MAYERNIK, 2020, *não paginado, tradução nossa*).

Os *value vocabularies* definem os valores dos elementos para descrição de um recurso. Eles não estabelecem informações de um recurso, mas sim conceitos relacionados a um recurso como pessoas, assuntos, idiomas, países etc. (ISAAC *et al.*, 2011). Como exemplos: o *Library of Congress Subject Headings*, o *Virtual International Authority File* (VIAF), a Classificação Decimal de Dewey (CDD) e o *GeoNames*.

Metadata element set é determinado como "[...] um conjunto de elementos de metadados que definem as classes e atributos utilizados para descrever entidades de interesse" (ISAAC *et al.*, 2011, *não paginado, tradução nossa*). O conjunto de elementos também designa um conjunto completo de elementos de metadados, assim como a codificação dos elementos e estrutura em uma linguagem de marcação (ZENG; QIN, 2008). Podem ser exemplificados: o *Dublin Core Metadata Terms*, os elementos do *Resource Description and Access* (RDA), do *Simple Knowledge Organization System* (SKOS) e o *Friend of a Friend vocabulary* (FOAF) (BAKER *et al.*, 2011).

Nesse sentido, a literatura aponta que, em muitos casos, dados e metadados são tratados como sinônimos. Segundo Wickett *et al.* (2013, *não paginado, tradução nossa*), "[...] os componentes de dados e metadados estão entrelaçados: nenhuma distinção estrutural permite uma discriminação imediata entre dados e metadados".

Exemplificando uma situação no contexto das bibliotecas, Jeffery *et al.* (2014, *não paginado, tradução nossa*) distinguem que "[...] para o pesquisador, o registro da biblioteca são metadados para descobrir um livro ou artigo de interesse. Para o bibliotecário, o registro pode ser utilizado como dados para analisar a completude relativa das coleções por assunto, por editora, por ano etc.". Ou seja, o mesmo objeto (dado/metadado) pode ser considerado metadado, como forma de descoberta e busca de um recurso informacional no catálogo, por exemplo.

Como também, poderá ser avaliado como dado, no momento em que o bibliotecário começar a analisar os registros bibliográficos contidos na biblioteca como um todo, para realizar análises do acervo, de empréstimos, entre outras possibilidades.

De acordo com Hyvönen (2012, p. 10, *tradução nossa*), "[...] em torno de 2005, as ideias sobre *Linked Data* e *Web* de dados começaram a ganhar impulso como uma abordagem simples para *Web* Semântica, focada na publicação de grandes conjuntos de dados existentes e usando apenas ontologias RDF simples e leves". Esse processo foi um dos movimentos para ressaltar a importância dos dados e a ampliação das pesquisas na área.

Além disso, os termos são essenciais nas discussões sobre a *Web* Semântica e para os princípios *Linked Data*. Os quatro princípios, definidos por Tim Berners-Lee em 2006, são sugestões de boas práticas para a estruturação de dados. No entanto, não são tidos como requerimentos obrigatórios. Esses princípios são:

1. Use URIs como nomes para as coisas;
2. Use HTTP URIs para que as pessoas possam procurar esses nomes;
3. Quando alguém procurar um URI, fornecer informações úteis, usando os padrões (RDF, SPARQL);
4. Incluir links para outros URIs, para que eles possam descobrir mais coisas.

No contexto dos metadados, Pomerantz (2015, p. 158, *tradução nossa*) faz uma releitura dos princípios do *Linked Data*. Dessa forma, Pomerantz (2015) enfatiza a questão da identificação dos recursos ao invés de coisas e ressalta a persistência dos metadados, conforme destacados nos princípios 3 e 4. Sendo que os princípios são redefinidos para:

1. Use URIs como identificadores para recursos;
2. Formate as URIs de acordo com o HTTP, para que os recursos possam ser encontrados facilmente, usando a tecnologia estabelecida;
3. Use padrões como o RDF para fornecer tanto o recurso e metadados sobre o recurso;
4. Fornecer *links* juntamente com esses metadados para outros URIs, para que mais recursos possam ser encontrados.

Paralelo a esse movimento, às tecnologias e aos novos paradigmas, tanto da Ciência como a *e-Science*, quanto da *Web* de documentos para *Web* de dados, impulsionou-se a importância dos dados para os processos do dia-a-dia. Nesse contexto, novos conceitos relacionados aos dados têm surgido, como a definição de *Big Data,* e mesmo as discussões de um campo específico para tratar de dados, como a Ciência de Dados.

Diante dos problemas e dos desafios de estudos sobre metadados na Ciência de Dados, Greenberg (2017) posiciona a importância dos metadados. De acordo com a autora, a Ciência de Dados é contextualizada como "[...] um campo interdisciplinar que possui como objetivo o estudo e avanço dos dados para obter *insights*. Uma empresa de Ciência de Dados pode permitir prever um fenômeno ou automatizar a tomada de decisões" (GREENBERG, 2017, p. 22, *tradução nossa*). A gama de dados encontrada pela Ciência de Dados concentra-se em diferentes formas. Greenberg (2017) destaca que são dados pequenos, grandes, estáticos, estruturados, não estruturados ou em fluxo contínuo e, ainda, são aplicadas metodologias científicas e estatísticas para o aprendizado dos dados.

No entanto, com relação aos metadados, Greenberg (2017) afirma que os esforços da Ciência de Dados dependem de uma descrição precisa dos dados, um subsídio necessário para os processos e para as aplicações. A autora também destaca que as estruturas de metadados garantem a confiabilidade dos dados e identificam uma infinidade de tópicos de pesquisa relacionados a metadados, muitos dos quais encontrados em relatórios governamentais e disciplinares similares em todo o mundo.

O uso e aplicação dos metadados e a importância dos cientistas de dados realizarem o tratamento da informação ou dos dados é destacada por Stanton *et al.* (2012) e por Curty e Serafim (2016). De acordo com Curty e Serafim (2016, p. 312), "[...] os cientistas de dados são responsáveis pela identificação, coleta, tratamento, transformação, análise, visualização e curadoria de grandes conjuntos de dados heterogêneos.". Esses aspectos foram descritos por Mishra e Chang (2015) e por Rautenberg e Carmo (2019. p. 62), que em seus trabalhos ressaltam a importância do gerenciamento, da proveniência, da curadoria e do arquivamento de dados e dos metadados, principalmente no âmbito do *Big Data*. Segundo Rautenberg e Carmo (2019, p. 62):

> [...] a Ciência da Informação contribui no fomento das competências de organização e representação de dados e informação, privilegiando os serviços de coleta, registro, filtragem, classificação e entrega de dados e seus metadados às atividades reservadas à camada da Ciência de Dados.

Os autores destacam as habilidades e a interdisciplinaridade da Ciência da Informação e Ciência da Computação. Tal relacionamento auxilia e possibilita aos gestores uma melhor visão para o processo de tomada de decisão guiada por dados, mesmo que seja uma grande proporção de dados, como no *Big Data*. Rautenberg e Carmo (2019, p. 62) salientam que os profissionais da Ciência da Informação e da Computação podem atuar juntos em diversas frentes da Ciência de Dados e do *Big Data*.

A relação entre Ciência da Informação e *Big Data* já foi relatada por diversos autores como: Dias e Vieira (2013), Souza, Ribeiro e Porto (2013), Ribeiro (2014), Araújo Júnior e Souza (2016), Dutra e Macedo (2016), Rodrigues, Dutra e Dias (2017), Coneglian, Gonçalez e Santarém Segundo (2017), Souza, Martins e Ramalho (2018), Andrade, Gonçalez, Berti Júnior, Baptista e Coneglian (2020), Reis e Sá (2020), dentre outros. Nesta relação, os autores realçam o termo dado ao tratamento e processamento de grande volume de dados.

Araújo Júnior e Souza (2016) relacionam as etapas do ciclo documentário (seleção, aquisição, registro ou tombamento, descrição, análise ou condensação, indexação, armazenamento dos documentos, armazenamento da representação, processamento da informação, produtos do processamento, interrogação e busca e recuperação da informação) proposto por Robredo (2005) como subsídio para a definição do tratamento dos dados no *Big Data*.

Corroborando com essa linha, outra questão que refulge é sobre as atividades da curadoria digital propostas por Higgins (2008), que incluem a descrição e representação, planejamento da preservação, acompanhamento e participação da comunidade, curadoria e preservação, conceitualização, criação e/ou recepção, avaliação e seleção, admissão, ações de preservação, armazenamento, acesso, uso e reuso, transformação, descarte, reavaliação e migração. Dutra e Macedo (2016) e Rautenberg e Carmo (2019) propõem que essas atividades sejam empregadas no contexto do *Big Data*.

Neste contexto, os metadados são fundamentais para possibilitar o gerenciamento de grande volume de dados. Triques, Arakaki e Castro (2020) discutiram e apresentaram elementos que relacionam a curadoria digital com a representação da informação, em especial, utilizando os metadados para auxiliar nas atividades curatoriais. Os metadados são evidenciados no contexto do *Big Data* por Reis e Sá (2020). De acordo com os autores,

> A proliferação de formatos de arquivos digitais e a explosão informacional vivida atualmente, com a produção de conteúdo digital por smartphones e outros dispositivos conectados à internet, geram uma grande quantidade de metadados. A quantidade disponível e sua relação indissociável do cotidiano do indivíduo tornam os metadados a mais rica fonte para a análise e extração de informações sobre consumidores e seus hábitos, despertando o interesse de empresas que buscam atingir vantagem competitiva através de ferramentas de análise de dados. Os metadados são, portanto, o principal insumo do *Big Data* (REIS; SÁ, 2020, p. 237).

Conforme observado, os metadados são a fonte primária para o *Big Data*, seguindo na perspectiva da interpretação documental (dados como metadados), conforme destacado por Furner (2016).

Nesta perspectiva interpretativa documental de Furner (2016), do que são os dados tidos como metadados, Magalhães *et al.* (2014) esclarecem que o *Big Data* é uma grande massa de dados com características estruturadas e não-estruturadas que busca a autenticidade dos objetos (dado/metadado) para dar sentido às informações que podem agregar alguma forma de valor a empresas e governos.

A partir dessa relação, os textos de Coneglian, Gonçalez e Santarém Segundo (2017, p. 138), e Reis e Sá (2020) apontam o papel do bibliotecário e do profissional da informação no *Big Data*, que pode facilitar e permitir a descoberta e recuperação dos dados, inferir na manutenção e qualidade dos dados, arquivamento, representação e preservação de dados, além de possibilitar a agregação de valor dos dados por meio dos metadados.

O termo também foi utilizado para fundamentar o *Big Data* e sua relação com os padrões de representação *Resource Description Framework* (RDF); *Simple Knowledge Organization System* (SKOS) e *Ontology Web Language* (OWL) pelos autores Souza, Martins e Ramalho (2018). Além disso, os autores relacionam esses padrões com os conceitos da *Web* Semântica e concluem que "[...] foi verificado que os modelos de representação analisados contribuem para interligar grandes volumes de dados sem perder o contexto no qual são originados, favorecendo um melhor entendimento do *Big Data* e os novos paradigmas de representação em ambientes digitais." (SOUZA; MARTINS; RAMALHO, 2018, p. 18).

Conforme visto, diversos autores relacionam a Ciência de Dados e o *Big Data* com as questões do tratamento e gerenciamento de metadados, utilizando-se de aportes teóricos e metodológicos, além de técnicas da Representação e Organização da Informação e do Conhecimento para estruturação dos dados.

CONSIDERAÇÕES FINAIS

Diante do levantamento e das discussões apresentadas no texto, observou-se que na literatura da área de Ciência da Informação há diversas relações entre os conceitos de dados e metadados. Como visto, os metadados podem ter diversas interpretações, como clássica, documental, eclesiástica, geométrica, matemática, epistêmica, informacional, computacional e diafórica. Em muitos casos, a depender da interpretação e do contexto que são tratados, os dados podem ser entendidos como metadados ou documento.

Na tentativa de esclarecer essas relações, em muitos casos, pode depender da perspectiva em que analisa os dados e de seu contexto. Ao analisar da ótica do usuário da biblioteca, o profissional da informação, bibliotecário, constrói registros informacionais, extrai e descreve os metadados referentes a algum recurso informacional para posterior busca pelo usuário. Assim, esta busca por uma informação a partir de metadados. Quando esse conjunto de registros criados são usados para a criação de relatórios ou pesquisa, pode-se considerar essas informações como dados, referindo-se ainda a esses registros como uma forma de documento e uma evidência dos recursos que constam na biblioteca, por exemplo.

O desenvolvimento tecnológico, atrelado às discussões e à formalização da *Web* Semântica, Ciência dos dados, *Big data*, entre outros conceitos, evidenciou a importância dos dados. Entretanto, verificou-se que há poucos estudos que contextualizam os metadados na Ciência dos Dados, e uma tendência de estudos sobre *Big Data*.

Ao final, não foi possível delimitar uma definição única dos dois conceitos, mas pôde-se verificar a diversidade epistemológica encontrada sobre a temática. Apesar disso, muitos autores discordam com esses conceitos apresentados e se destaca a necessidade de aprofundamento em trabalhos futuros, com a finalidade de fomentar discussões sobre a temática e contextualizá-la. Além disso, ainda é necessário delinear posicionamentos e perspectivas divergentes entre pesquisas.

REFERÊNCIAS

ANDRADE, M. C. *et al.* Ciência responsável dos dados: imparcialidade, precisão, confidencialidade, e transparência dos dados. *Informação & Informação*, v. 25, n. 2, p. 26-48, 2020. Disponível em: http://www.uel.br/revistas/uel//index.php/informacao/article/view/38467. Acesso em: 25 set. 2020.

ALVES, R. C. V. *Metadados como elementos do processo de catalogação.* 2010. 132p. Tese (doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista Júlio Mesquita Filho, Marília, 2010. Disponível em: http://repositorio.unesp.br/handle/11449/103361. Acesso em: 8 set. 2019.

ALVES, R. C. V.; SANTOS, P. L. V. A. C. *Metadados no domínio bibliográfico*. Rio de Janeiro: Intertexto, 2013.

ARAKAKI, F. A. *Metadados administrativos e a proveniência dos dados*: modelo baseado na família PROV. 2019. 139p. Tese (doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista Júlio Mesquita Filho, Marília, 2019. Disponível em: https://repositorio.unesp.br/handle/11449/180490. Acesso em: 8 set. 2019.

ARAÚJO JÚNIOR, R. H.; SOUSA, R. T. B. Estudo do ecossistema de big data para conciliação das demandas de acesso, por meio da representação e organização da informação. *Ciência da Informação*, v. 45, n. 3, 2016. Disponível em: http://revista.ibict.br/ciinf/article/view/4057. Acesso em: 25 set. 2020.

BAKER, T. *et al. Library Linked Data Incubator Group Final Report.* W3C Incubator Group Report, 2011. Disponível em: http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/. Acesso em: 8 set. 2019.

BERNERS-LEE, T. *Linked Data:* Design Issues. [s.l.]: W3C, 2006. Disponível em: http://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 25 jun. 2015.

BUCKLAND, M. *Information and Society*. Cambridge: MIT Press, 2017.

CONEGLIAN, C. S.; GONÇALEZ, P. R. V. A.; SANTARÉM SEGUNDO, J. E. O profissional da informação na era do big data. *Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação*, v. 22, n. 50, p. 128-143, 2017. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2017v22n50p128. Acesso em: 12 mar. 2021.

CURTY, R. G.; SERAFIM, J. da S. A formação em ciência de dados: uma análise preliminar do panorama estadunidense. *Informação & Informação*, v. 21, n. 2, p. 307-331, dez. 2016. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/27945. Acesso em: 24 set. 2020.

DIAS, G. A.; VIEIRA, A. A. N. Big data: questões éticas e legais emergentes. *Ciência da Informação*, v. 42, n. 2, 2013. Disponível em: http://revista.ibict.br/ciinf/article/view/1380. Acesso em: 12 mar. 2021.

DUTRA, M. L.; MACEDO, D. D. J. Curadoria digital: proposta de um modelo para curadoria digital em ambientes big data baseado numa abordagem semi-automática para a seleção de objetos digitais. *Informação & Informação*, v. 21, n. 2, p. 143-169, 2016. Disponível em: http://www.uel.br/revistas/uel//index.php/informacao/article/view/27176. Acesso em: 12 mar. 2021.

EDWARDS, P. *et al.* Science Friction: Data, Metadata, and Collaboration. *Social Studies of Science, v.* 41, n. 5, p. 677-690, 2011.

FURNER J. "Data": The data. In: KELLY, M.; BIELBY J. (Eds.) *Information Cultures in the Digital Age*. Wiesbaden: Springer VS, 2016.

FURNER, J. Definitions of "Metadata": A Brief Survey of International Standards. *Journal of the Association for Information Science and Technology*, 2019.

FURNER, J. Philosophy of data: why?. *Education For Information*, v. 33, n. 1, p. 55-70, 2017.

GILLILAND, A. J. Setting the Stage. In: BACA, M. (Org.). *Introduction to Metadata.* 3. ed. Los Angeles: Getty Research Institute, 2016. Disponível em: http://www.getty.edu/publications/intrometadata/. Acesso em: 8 set. 2019.

GREENBERG, J. Big Metadata, Smart Metadata and Metadata Capital: toward greater synergy between data science and metadata. *Journal of Data and Information Science*, v. 2, n. 3, p. 19-36, 2017.

HAYNES, D. *Metadata for information management and retrieval.* [s.l.]: Facet Publishing, 2004.

HIGGINS, S. The DCC Curation Lifecycle Model. *The International Journal of Digital Curation*, n. 1, v. 3, 2008.

HYVÖNEN, E. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. EUA: Morgan & Claypool Publishers, 2012.

ISAAC, A. *et al. Library Linked Data Incubator Group:* Datasets, Value Vocabularies, and Metadata Element Sets: W3C Incubator Group Report 25 October 2011. W3C, 2011. Disponível em: http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset20111025/.

JEFFERY, K. *et al.* A 3-Layer Model for Metadata. INTERNATIONAL CONFERENCE ON DUBLIN CORE AND METADATA APPLICATION, 13., Portugal, *Anais*… DCMI, EUA. 2014. Disponível em: http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/199/199. Acesso em: 8 set. 2019.

JOUDREY, D. N.; TAYLOR, A. G.; WISSER, K. M. *The organization of information.* 4. ed. California: Libraries Unlimited, 2018.

MAGALHAES, V. R. V. *et al.* O uso do Big Data na violação da privacidade dos usuários para estratégias de negócios. In: ENCONTRO REGIONAL DE COMPUTAÇÃO E SISTEMAS DE INFORMAÇÃO, 3.; ENCONTRO REGIONAL DE COMPUTAÇÃO E SISTEMAS DE INFORMAÇÃO, *Anais*..., Manaus, 2014.

MAYERNIK, M. Metadata Accounts: Achieving Data and Evidence in Scientific Research. *Social Studies of Science, v.* 49, n. 5, p. 732-757, 2019.

MAYERNIK, M. Metadata**.** In: HJØRLAND, B.; GNOLI, C. *Encyclopedia of Knowledge Organization.* ISKO, 2020. Disponível em: https://www.isko.org/cyclo/metadata. Acesso em: 30 abr. 2020.

MÉNDEZ RODRÍGUEZ, E. *Metadatos y recuperación de información*. Gijón: Ediciones Trea, 2002.

PETERSON. D. *Forms of representation:* an interdisciplinary theme for cognitive science. Wiltshire: Cromwell Press, 1996.

POMERANTZ, J. *Metadata.* Cambridge: The MIT Press, 2015.

RAUTENBERG, S.; CARMO, P. R. Big data e ciência de dados. *Brazilian Journal of Information Science*, v. 13, n. 1, p. 56-67, 2019.

REIS, L. C. R.; Sá, M. I. F. E. Big data: um novo campo de atuação para bibliotecários. *Prisma.com*, Portugal, n. 41, p. 231-250, 2020.

RIBEIRO, C. J. S. Big data: os novos desafios para o profissional da informação. *Informação & Tecnologia*, v. 1, n. 1, p. 96-105, 2014.

ROBREDO, J. *Documentação de hoje e de amanhã*. Brasília: Edição de autor, 2005.

RODRIGUES, A. A.; DUARTE, E. N.; DIAS, G. A. Desafios da gestão de dados na era do big data: perspectivas profissionais. *Informação & Tecnologia*, v. 4, n. 2, p. 63-79, 2017.

SANTOS, P. L. V. A. da C.; SANTANA, R. C. G. Dado e Granularidade na perspectiva da Informação e Tecnologia: uma interpretação pela Ciência da Informação. *Ciência da Informação*, v. 42, n. 2, jan. 2013. Disponível em: http://revista.ibict.br/index.php/ciinf/article/view/228. Acesso em: 8 set. 2019.

SANTOS, P. L. V. A. C.; SIMIONATO, A. C.; ARAKAKI, F. A. Definição de metadados para recursos informacionais: apresentação da metodologia BEAM. *Informação & Informação*, Londrina, v. 19, n. 1, p. 146-163, fev. 2014. Disponível em: http://repositorio.unesp.br/handle/11449/114736. Acesso em: 8 set. 2019.

SANTOS, P. L. A. C.; VIDOTTI, S. A. B. G. Perspectivismo e tecnologias de informação e comunicação: acréscimos à Ciência da Informação. *DataGramaZero: revista de Ciência da Informação*, Rio de Janeiro, v. 10, n. 3, 2009.

SOUZA, J. L.; MARTINS, P. G. M.; RAMALHO, R. A. S. Modelos de representação semântica na era do big data. *Brazilian Journal of Information Science*, v. 12, n. 3, p. 34-40, 2018.

STANTON, J. *et al.* Interdisciplinary data scienceeducation. In: XIAO, N.; MCEWEN, L. R. *SpecialIssues in Data Management.* Washington: American ChemicalSociety, 2012. p. 97-113.

TRIQUES, M. L.; ARAKAKI, A. C. S.; CASTRO, F. F. Aspectos da representação da informação na curadoria digital. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, Florianópolis, v. 25, p. 1-21, maio 2020. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2020.e69898. Acesso em: 25 set. 2020.

WICKETT, K. M. *et al.* Identifying content and levels of representation in scientific data. *Proceedings of The American Society For Information Science And Technology,* v. 49, n. 1, p. 1-10, 2013.

ZENG, M. L. QIN, J. *Metadata.* New York: Neal-Schuman Publishers, 2008.

ZENG, M. L. QIN, J. *Metadata.* 2. ed. London: facet publishing, 2016.

AGRADECIMENTOS

Agradecemos ao Grupo de Pesquisa "Dados e Metadados" e o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), processo 431612/2016-1.

Modelagem de metadados multimídia: uma proposta ontológica baseada em reúso

Daniela Lucas da Silva Lemos
Doutorado em Ciência da Informação, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, MG, Brasil. Professora do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil.
http://lattes.cnpq.br/9280443047358807
http://www.biblioteconomia.ufes.br/daniela-lucas-da-silva-lemos
https://orcid.org/0000-0003-1565-7366
E-mail: danielalucas@hotmail.com

Submetido em:15/09/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021.

RESUMO

Nos últimos anos, observou-se um crescimento significativo de dados semanticamente relacionados e distribuídos na Web. Nesse contexto, padrões de metadados recomendados pelo *World Wide Web Consortium* vêm sendo utilizados para descrever e representar recursos multimídia, possibilitando ampliar os pontos de acesso e melhorar a gestão, a organização e a recuperação de recursos digitais na rede. Um problema comumente verificado nas bases de dados institucionais está no tratamento integrado de dados heterogêneos e na ausência de padronização nos formatos de descrição. A descrição de inúmeros itens, geralmente, é realizada de maneira independente, com padrões idiossincráticos de descrição, ressaltando diferentes características a serem descritas e diferentes terminologias para descrevê-las, desconsiderando requisitos de interoperabilidade entre as comunidades. A presente pesquisa buscou avançar nas formas de representação de documentos multimídia com textos, vídeos, imagens, modelos 3D, áudios, propondo um Modelo Ontológico de Referência Multimídia capaz de organizar semanticamente tipologias de metadados para descrição de conteúdo multimídia diante de variados contextos e necessidades. Metodologicamente, a proposta de modelagem foi fundamentada em ontologias multimídia mais bem colocadas em um *ranking* obtido a partir do guia *NeOn Methodology*, o que assegurou a seleção de recursos de conhecimento adequados diante de requisitos funcionais e não funcionais determinados. O Modelo Ontológico de Referência Multimídia propôs classes e relacionamentos ontológicos fundamentais provenientes de mapeamentos e alinhamentos semânticos entre as ontologias reutilizadas, promovendo uma arquitetura abrangente para a organização semântica de metadados multimídia endereçados, principalmente, a aplicações que lidam com recursos de informação na Web.

Palavras-chave: Recursos multimídia. Anotação multimídia. Modelo ontológico de referência multimídia. Ontologias. Interoperabilidade.

Multimedia metadata modeling: an ontological proposal based on reuse

ABSTRACT

In recent years, there is a significant growth of semantically related and distributed data on the Web. In this context, metadata standards recommended by the World Wide Web Consortium have been used to describe and represent multimedia resources, making it possible to expand access points and improve the management, organization and recovery of digital resources on the network. A problem commonly found in institutional databases is the integrated treatment of heterogeneous data and the lack of standardization in description formats. The description of numerous items is generally carried out independently, with idiosyncratic standards of description, emphasizing different characteristics to be described and different terminologies to describe them, disregarding interoperability requirements between communities. The present research sought to advance in the forms of representation of multimedia documents with texts, videos, images, 3D models, audios, proposing an Ontological Model of Multimedia Reference capable of semantically organizing types of metadata to describe multimedia content in different contexts and needs. Methodologically, the modeling proposal was based on multimedia ontologies better placed in a ranking obtained from the NeOn Methodology guide, which ensured the selection of adequate knowledge resources in light of functional and non-functional requirements. The Ontological Model of Multimedia Reference proposed fundamental ontological classes and relationships from mappings and semantic alignments between reused ontologies, promoting a comprehensive architecture for the semantic organization of multimedia metadata addressed mainly to applications that deal with information resources on the Web.

Keywords*: Multimedia resources. Multimedia annotation. Ontological Model of Multimedia Reference. Ontologies. Interoperability.*

Modelado de metadatos multimedia: una propuesta ontológica basada en la reutilización

RESUMEN

En los últimos años, ha habido un crecimiento significativo en los datos semánticamente relacionados y distribuidos en la Web. En este contexto, estándares de metadatos recomendados por el World Wide Web Consortium se han utilizado para describir y representar recursos multimedia, lo que permite ampliar los puntos de acceso y mejorar gestión, organización y recuperación de recursos digitales. Un problema que se encuentra comúnmente en las bases de datos institucionales es el tratamiento integrado de datos heterogéneos y la falta de estandarización en los formatos de descripción. La descripción de numerosos elementos generalmente se lleva a cabo de forma independiente, con estándares de descripción idiosincráticos, enfatizando diferentes características para ser descritas y diferentes terminologías para describirlas, sin tener en cuenta los requisitos de interoperabilidad. La presente investigación buscó avanzar en las formas de representación de documentos multimedia, proponiendo un Modelo Ontológico de Referencia Multimedia capaz de organizar semánticamente tipos de metadatos para describir contenido multimedia en diferentes contextos y necesidades. Metodológicamente, la propuesta de modelado se basó en ontologías multimedia mejor ubicadas en un ranking obtenido de la guía de Metodología NeOn, que aseguró la selección de recursos de conocimiento adecuados a la luz de requisitos funcionales y no funcionales. El Modelo Ontológico de Referencia Multimedia propuso clases y relaciones ontológicos fundamentales a partir de mapeos y alineamientos semánticos entre ontologías reutilizadas, promoviendo una arquitectura para la organización semántica de metadatos multimedia dirigidos principalmente a aplicaciones que manejan recursos de información en la Web.

Palabras clave*: Recursos multimedia. Anotación multimedia. Modelo Ontológico de Referencia Multimedia. Ontologías. Interoperabilidad.*

INTRODUÇÃO

Nos últimos anos, observou-se um crescimento significativo de dados semanticamente relacionados e distribuídos na Web. Nesse contexto, padrões de metadados recomendados pelo *World Wide Web Consortium* (W3C) vêm sendo utilizados para descrever e representar recursos multimídia, possibilitando ampliar os pontos de acesso e melhorar a gestão, a organização e a recuperação de recursos digitais na rede. Entretanto, o relacionamento entre multimídia e Web de dados ainda é um ramo de pesquisa que carece de estudos avançados, voltados a tecnologias eficientes para geração, exposição, descobrimento e consumo de recursos multimídia semanticamente vinculados na Web (SCHANDL *et al.*, 2012; SILVA; SOUZA, 2014; LEMOS; SOUZA, 2020).

Um recurso multimídia contempla um documento composto, que faz referência a vários tipos de objetos, tais como vídeo, texto, som, imagem, modelo tridimensional (3D), dentre outros, e ainda pode ser dividido em partes que resultam em tipos de mídias específicos (SCHANDL *et al.*, 2012). Adjeroh e Nwosu (1997) acrescentam ainda que alguns tipos de dados multimídia, como vídeo, áudio e sequências de animação, possuem requisitos temporais que implicam diretamente na representação, no armazenamento, na transmissão, manipulação e apresentação do dado. De forma similar, imagens, modelos 3D e vídeos possuem restrições espaciais em seus conteúdos, concernentes a relações espaciais entre objetos individuais pertencentes a uma imagem, a uma réplica digital em 3D ou a um quadro de vídeo, respectivamente. Assim, na representação de recursos multimídia, deve-se considerar características particulares, incluindo relações espaciais entre elementos de interesse no conteúdo e relações temporais na ocorrência de eventos em dado período de tempo.

Atualmente, recursos multimídia tornam-se onipresentes no lazer, no aprendizado, nas artes, na comunicação, no comércio, nas ciências, tomando formatos de arquivos digitais produzidos e disponibilizados, geralmente, em repositórios digitais na Web.

Exemplos disso, seriam as coleções digitais organizadas em bases de dados oriundas de museus e outras instituições de cultura responsáveis pela guarda e divulgação de obras de arte e documentos históricos (HYVÖNEN, 2012). Nesse cenário, observa-se, principalmente, um aumento de práticas de digitalização de acervos institucionais e a inserção de itens e coleções em sistemas de informação, como os repositórios digitais, que promovem condições necessárias a novas formas de organização, acesso e recuperação da informação na rede (POTENZIANI *et al.*, 2018), culminando no surgimento de grandes bases de dados de objetos digitais heterogêneos.

Assim, documentos institucionais digitalizados ou produzidos em formato digital podem estar disponibilizados nessas bases de dados na forma de texto, imagem, modelo 3D, áudio e vídeo, isoladamente ou em conjunto, necessitando de métodos e técnicas específicos para curadoria (processos de avaliação, de adição de valor, reformatação, agregação e reúso de dados) e disseminação de recursos digitais à sociedade em rede. Para tanto, torna-se necessário que esses recursos sejam descritos em seus aspectos de mídia e de conteúdo para obtenção de uma documentação condizente com a realidade do domínio, em termos de acesso e recuperação de informações.

Além dos aspectos de curadoria adequada a dados multimídia, o acesso e o consumo de coleções de objetos na rede vislumbram, muitas vezes, a necessidade de integração semântica e disponibilização global, a fim de possibilitar o compartilhamento, a interligação entre vários acervos e o reúso de conteúdos digitais relevantes aos provedores de conteúdo e seus usuários finais. Contudo, constata-se que bases de dados institucionais habitualmente utilizam padrões idiossincráticos de descrição, ressaltando diferentes características a serem representadas em detalhes e diferentes terminologias para descrever seus recursos de informação a partir de uma exigência da própria comunidade e desconsiderando requisitos de interoperabilidade.

Esta pode ser compreendida como a capacidade de diversos sistemas e organizações trabalharem em conjunto para garantir que pessoas, organizações e sistemas computacionais interajam na troca de informações de maneira eficaz e eficiente (LEMOS; MENDONÇA; SOUZA, 2020). Trata-se de uma das principais metas atuais do consórcio W3C, que vem se dedicando ao desenvolvimento de padrões para avançar nessa perspectiva. Assim, grandes desafios são lançados, especialmente no que se refere à integração de bases de dados heterogêneas e sistemas institucionais que já possuem acervos digitalizados na rede, demandando estratégias de busca de soluções inteligentes para a descrição de seus recursos de informação multimídia na Web.

Pesquisas em Ciência da Informação e Ciência da Computação (BERNERS-LEE; HENDLER; LASSILA, 2001; GUIZZARDI, 2005; SILVA, 2014) têm concentrado seus esforços para empreender melhorias nos sistemas de recuperação da informação, sobretudo quanto à organização e representação da informação, objetivando uma Web de conteúdo semântico e interoperável. Nesse sentido, evidenciam-se estudos sobre padrões de metadados, modelos conceituais, vocabulários controlados e ontologias para o tratamento descritivo e temático de documentos em várias mídias, visando a necessidade de integração semântica e disponibilização global de recursos de informação em rede.

Durante as últimas décadas, surgiram várias iniciativas na produção de ontologias baseadas em *Resource Description Framework* (RDF) e *Ontology Web Language* (OWL) voltadas à descrição dos dados multimídia (LEMOS; SOUZA, 2019), cujos esforços objetivaram transformar padrões de metadados multimídia, como o MPEG-7 *International Organization for Standardization/International Electrotechnical Commission* (ISO/IEC) (SALEMBIER; SMITH, 2001), em formatos semelhantes a ontologias. O padrão é usado para prover um vocabulário rico e comum a recursos multimídia, incluindo descritores primitivos, extraídos da própria mídia e, de alto nível, destinados à descrição semântica de conteúdo da mídia.

Nesta pesquisa, ontologias são vistas como modelos de anotação (LEMOS; SOUZA, 2020) em uma perspectiva de tratamento semântico destinado a dados e metadados envolvidos no processo de representação, o que permite descrever e interligar recursos por meio de qualificadores, incluindo conceitos, instâncias, propriedades e restrições, cujas proposições são asseguradas pela definição de axiomas. O modelo é indicado à anotação semântica de documentos, que Shadbolt, Berners-Lee e Hall (2006) esclarecem ser uma abordagem subjacente aos conceitos preconizados pela Web Semântica, juntamente com sua proposta de dados abertos interligados (do inglês, *Linked Open Data* - LOD) (MACHADO; SOUZA; SIMÕES, 2019), no que tange ao fornecimento de significado à organização da informação por meio de conexões lógicas entre os dados.

Considera-se, portanto, que esforços na construção de ontologias podem ser poupados, tendo em vista a exploração de vocabulários existentes e disponíveis para reúso em comunidades de interesse. A prática de reúso é reconhecida como sendo um importante passo na construção de vocabulários semânticos, incluindo ontologias. Com o reúso de recursos já existentes poupa-se tempo e esforço, ao invés de se começar a construção do zero (SUÁREZ-FIGUEROA; GÓMEZ-PÉREZ; FERNÁNDEZ-LÓPEZ, 2012). Contudo, surgem desafios na identificação e seleção de uma variedade de vocabulários e ontologias disponíveis e que precisam ser compatíveis com as entidades reais de um domínio específico. Alguns desses desafios seriam: a disponibilização de fontes documentais adequadas, envolvendo as conceituações de ontologias, úteis especialmente aos processos de reúso; os alinhamentos e mapeamentos semânticos, visando à consistência entre os elementos cotejados; a modelagem conceitual condizente com a realidade do domínio; os aspecto de usabilidade na interação com o *software* e os dados; e a manutenção das ontologias envolvidas em processos de reúso.

Abrem-se, desse modo, oportunidades e desafios, para a presente pesquisa, de responder às seguintes questões: *i)* como expressar a estrutura conceitual subjacente à descrição da realidade documental de tipo multimídia? *ii)* qual a estratégia para selecionar e alinhar eficientemente vocabulários e ontologias multimídia concebidos por comunidades distintas para cobrir satisfatoriamente aspectos ontológicos da realidade documental de tipo multimídia? e *iii)* como organizar sistematicamente tipologias de metadados existentes para descrição de conteúdo multimídia diante de variados contextos e necessidades?

Na proposição do preenchimento dessas lacunas, este artigo objetiva apresentar uma proposta de modelagem ontológica de referência multimídia fundamentada em recursos de conhecimento representativos e selecionados por meio de um *ranking* de ontologias candidatas a reúso. O sentido de "referência" é o que caracteriza o modelo como um artefato subjacente a esforços multidisciplinares de pesquisas voltados a modelos e tecnologias para processamento de metadados que se ocupam da anotação multimídia.

As contribuições desta pesquisa vão desde trazer à luz potenciais vocabulários para o domínio de anotação multimídia, incluindo padrões de metadados, ontologias e modelos conceituais subjacentes, até uma proposta bem fundamentada de um modelo conceitual ontológico de descrição de recursos multimídia para domínios diversos em processo de publicação e consumo de seus dados na Web.

METODOLOGIA

A presente pesquisa foi classificada como sendo de natureza qualitativa e quantitativa, em busca do entendimento do que estava por trás do fenômeno investigado, a saber, as *formas de representaçã*o *de recursos multimídia em rede*. Para o aspecto quantitativo, houve a necessidade do uso de formatos numéricos para mensurações de critérios avaliativos diante de análises de ontologias, que foram pontuadas e classificadas por métodos estatísticos. Também foi realizado um estudo exploratório, descritivo e explicativo à luz de literatura científica e material empírico específico, o que tornou esta pesquisa bibliográfica e documental.

Para a proposta de modelagem ontológica de referência multimídia, foi necessária a adoção de um guia metodológico atual, testado e validado em diferentes domínios e áreas, e, ainda, que seguisse diretrizes para a construção de ontologias em rede *Linked Open Data*. Para tanto, realizou-se uma revisão na literatura da área de Engenharia de Ontologias, e, em meio a um conjunto de propostas aventadas (SILVA; SOUZA; ALMEIDA, 2008; SUÁREZ-FIGUEROA; GÓMEZ-PÉREZ; FERNÁNDEZ-LÓPEZ, 2012), selecionou-se o guia *NeOn Methodology*, por dispor de práticas em iniciativas LOD e ser oriundo de *frameworks* metodológicos amplamente aceitos em áreas maduras, como Engenharia de Software e Engenharia do Conhecimento.

A metodologia de engenharia de ontologias *NeOn* abrange nove cenários que sugerem uma série de passos flexíveis para o desenvolvimento de ontologias. Os cenários envolvidos cobrem situações em que ontologias disponíveis em repositórios da Web necessitem, por exemplo, de reengenharia, alinhamento, modularização, localização, suporte em diferentes línguas e culturas, integração com padrões de projeto e recursos não ontológicos, tais como padrões de metadados, tesauros, taxonomias, dentre outros. Dos nove cenários indicados no guia, seis foram selecionados (cenários 1, 2, 3, 5, 6 e 8) a partir de uma adequação de fases de modelagem propositivas na presente pesquisa, conforme abaixo:

a) Identificação e seleção de recursos ontológicos e não ontológicos no domínio de anotação multimídia: cenário 2 e 3;

b) Análiser e comparação de ontologias multimídia à luz de requisitos previamente propostos: cenário 3;

c) Seleção de ontologias multimídia adequadas ao reúso de recursos de conhecimento destinados à construção do modelo proposto: cenários 3 e 5;

d) Desenvolvimento de um modelo conceitual baseado em ontologias para o domínio de anotação multimídia: cenários 1, 6 e 8;

Os outros três cenários (4, 7 e 9) não foram considerados na proposição do modelo, porque o cenário 4 indica reengenharia de recursos ontológicos de forma individualizada (por ontologia), isto é, não prevê a reengenharia de um conjunto de recursos ontológicos pós-processo de integração entre eles. O cenário 6 mostrou-se mais adequado à realidade de construção do modelo, por indicar reengenharia após a criação de um novo recurso ontológico. O cenário 7 revela reúso de padrões de projetos de ontologias (do inglês, *Ontology Design Patterns*) disponíveis em repositórios específicos, ou seja, meta-modelos com melhores práticas de desenvolvimento de ontologias em contextos específicos, de modo a auxiliar o ontologista em soluções de modelagem. Apesar de o modelo fazer uso de padrões de projetos (com extensões multimídia) advindos de ontologias de alto nível, não se praticou, a *priori*, busca e reúso de soluções para padrões específicos de modelagem. Finalmente, o cenário 9 compreende casos em que o modelo precisa ser desenvolvido em diferentes linguagens para aplicações multilíngues, não comportando, ainda, características da proposição de modelagem desta investigação.

A partir da determinação dos cenários metodológicos para a presente pesquisa, a primeira subseção descreve os procedimentos metodológicos conduzidos para a identificação, seleção e análise das ontologias para anotação multimídia. Por conseguinte, a segunda subseção descreve os cenários metodológicos empregados na proposição do Modelo Ontológico de Referência Multimídia.

IDENTIFICAÇÃO, SELEÇÃO E ANÁLISE DE ONTOLOGIAS PARA ANOTAÇÃO MULTIMÍDIA

A primeira fase desta investigação foi procedida por uma análise de domínio dotada de fontes documentais, incluindo normas, artigos e bibliotecas de esquemas *Extensible Markup Language* (XML) relacionados a padrões para descrição ou anotação de recursos multimídia na Web de dados. O cenário 2 - *reúso e reengenharia de recursos não ontológicos* foi operacionalizado nessa fase da pesquisa, partindo da obtenção de um conjunto de elementos de parâmetro (SILVA, 2014) fundamentados nos padrões de metadados para descrição de recursos digitais ou multimídia, o MPEG-7 e o Dublin Core. As revisões na literatura, vale ressaltar, evidenciaram que grande parte de ontologias para anotação multimídia é construída seguindo esses padrões de metadados (SILVA; SOUZA, 2014).

Os elementos de parâmetro foram, então, definidos e organizados com base em categorias de metadados multimídia determinadas na atividade de aquisição de conhecimento sobre o domínio, a saber: metadados independentes de conteúdo, metadados dependentes de conteúdo e metadados descritivos de conteúdo. A primeira categoria foi organizada para metadados relacionados à gestão da mídia, incluindo produção, classificação, gestão de direitos a uso e informações técnicas. A segunda categoria foi organizada para metadados de nível baixo ou primitivos, incluindo aspectos visuais (cor, textura e forma), de movimento, de localização e spaço-temporal no conteúdo da mídia, além de aspectos envolvendo processamento de sinais de áudio. A terceira categoria foi organizada para metadados voltados a segmentos de mídia, à anotação de conteúdo semântico (descrição de evento, objeto, agente, lugar e tempo), navegação e acesso, organização de conteúdo em coleções, preferências de uso em relação a conteúdo, e a características de áudio de nível alto, incluindo descritores para tratamento de conteúdo falado.

Desse modo, os elementos de parâmetro foram considerados produto-base da atividade de aquisição de conhecimento sobre o domínio, servindo, portanto, para identificar, analisar e comparar ontologias para anotação multimídia no aspecto de características concernentes a padrões de metadados consolidados nas comunidades de Biblioteca Digital, Web Semântica e Multimídia.

A segunda e a terceira fase foram operacionalizadas pelo cenário 3 - *reúso de recursos ontológicos*, que compreende o reúso de possíveis recursos ontológicos existentes para construção ou aprimoramento de uma rede de ontologias. Os passos metodológicos seguidos foram: *i)* busca por ontologias candidatas a reúso; e *ii)* análise comparativa dos recursos ontológicos selecionados em *(i)*, a partir de critérios pré-definidos na pesquisa. Buscou-se, então, partindo das orientações do guia, identificar ontologias para anotação multimídia fazendo um levantamento na literatura e buscas em repositórios da Web Semântica. Após um processo de refinamento diante das ontologias previamente selecionadas para análise, nove ontologias foram escolhidas (LEMOS; SOUZA, 2019), a saber: Media Ontology; M3 Multimedia; *Multimedia Metadata Ontology* (M3O); *Bootstrapping Ontology Evolution with Multimedia Information* (Boemie); *Core Ontology for Multimedia* (COMM); Polysema MPEG-7 MDS; MPEG-7 de Hunter; SmartWeb; e Rhizomik.

A terceira e última fase foi conduzida por uma análise específica e comparativa das ontologias selecionadas, em que seus conteúdos (códigos) e documentações subjacentes foram inspecionados e examinados. Os 17 (dezessete) critérios determinados para analisar e avaliar as ontologias foram, em sua maioria, oriundos do guia, os quais se originaram de casos de uso em diversas experiências de projeto, envolvendo desenvolvimento e reúso de ontologias.

A organização desses critérios ocorreu em quatro dimensões relacionadas ao reúso de ontologias, elucidadas como segue: *i)* esforço de reúso dos recursos: estimativa de custos relacionados ao tempo e à economia necessários ao reúso da ontologia avaliada; *ii)* esforço de entendimento dos recursos: estimativa de esforços necessários ao entendimento do conteúdo da ontologia avaliada; *iii)* esforço de integração dos recursos: estimativa de esforços empreendidos com vistas à integração da ontologia avaliada à ontologia em desenvolvimento; e *iv)* confiabilidade dos recursos: análise da confiança com relação à ontologia avaliada diante de aspectos de tratamento semântico nas declarações (ex. axiomas presentes, recursos de conhecimento utilizados – padrões de metadados –, ontologias de fundamentação), avaliação (ex. testes disponíveis) e projetos renomados de que fazem uso.

O método para obtenção de pontuações para cada ontologia deu-se por média ponderada, abarcando pesos determinados e valores mensurados para os critérios (LEMOS; SOUZA, 2020). Para estes últimos, a escala de valores determinada foi de 0 a 3, correspondendo, na sequência, aos qualificadores (D)esconhecido, (B)aixo, (M)édio e (A)lto. Desse modo, a pontuação resultante para cada ontologia avaliada permaneceu sempre numa escala de 0 a 3, sendo que, para a dimensão relacionada ao esforço de reúso dos recursos, quanto menor o qualificador aplicado ao critério, melhor a avaliação da ontologia, tendo em vista critérios com influência negativa, como custo econômico e tempo requerido na aquisição da ontologia; já para as dimensões esforço de entendimento dos recursos, esforço de integração dos recursos e confiabilidade dos recursos, quanto maior o qualificador, melhor a avaliação da ontologia, considerando critérios com influência positiva que qualificam a ontologia nas respectivas dimensões. Assim sendo, o guia atribui os símbolos (-) e (+) aos pesos para se considerar tais influências na pontuação do *ranking* resultante.

PROPOSIÇÃO DO MODELO ONTOLÓGICO DE REFERÊNCIA MULTIMÍDIA

A proposta do Modelo Ontológico de Referência Multimídia foi subsidiada pelo conhecimento obtido nas análises específicas e comparativas realizadas e pelos recursos de conhecimento ontológicos (ontologias multimídia) e não ontológicos (padrões de metadados multimídia) envolvidos na pesquisa. As ontologias melhor classificadas foram novamente analisadas em suas estruturas de conhecimento a fim de se decidir quais recursos seriam selecionados para reúso.

As decisões de seleção foram determinadas diante de requisitos funcionais e não funcionais, elaborados com base nas constatações resultantes da análise comparativa envolvendo as ontologias candidatas a reúso. Requisitos funcionais são o conjunto de características que reflete a estrutura de conhecimento multimídia da ontologia; já os requisitos não funcionais são associados ao uso da ontologia em termos de usabilidade, confiabilidade, segurança, disponibilidade, manutenibilidade e tecnologias envolvidas. Assim, os cenários 1, 3, 5, 6 e 8 do guia metodológico NeOn são descritos a seguir na forma como foram usados na fase de desenvolvimento da modelagem conceitual.

O cenário 1 - *da especificação à implementação* foi utilizado especialmente para a elaboração da especificação de requisitos entre os quais o modelo conceitual deveria ser capaz de representar, de modo a identificar possíveis problemas que a ontologia deveria conseguir resolver. Os requisitos denominados funcionais foram descritos a partir das categorias de tipos de metadados multimídia determinados na atividade de aquisição de conhecimento, elucidada na subseção anterior. Ademais, o propósito e o escopo do modelo também foram incluídos na especificação do Modelo Ontológico de Referência Multimídia.

Os cenários 3 e 5 - *reúso, alinhamento e combinação de recursos ontológicos* foram considerados quando se definiu o modo de reúso, ou seja, de que forma os recursos ontológicos selecionados seriam reusados. O guia indica três modos, a saber: *i)* tais recursos são reusados conforme estão, ou seja, sem modificações; *ii)* eles devem passar por um processo de reengenharia, caso necessitem de mudanças (modo incluso no cenário 6); e *iii)* alguns recursos ontológicos devem ser combinados a fim de se obter um novo recurso (modo incluso no cenário 5).

O cenário 5 contemplou um conjunto de mapeamentos que abarca os elementos de interesse das ontologias analisadas (classes, propriedades e axiomas). Esse tipo de informação tornou-se útil à modelagem conceitual quando se necessitou estabelecer arranjos dos recursos de conhecimento ontológicos em grupos coerentes com os requisitos funcionais traçados na pesquisa para, finalmente, estabelecer os mapeamentos semânticos demandados na atividade de modelagem.

O cenário 6 - *reúso, alinhamento, combinação e reengenharia de recursos ontológicos* foi usado na medida em que alguns recursos de conhecimento advindos de ontologias selecionadas para reúso precisaram de mudanças para se ajustarem aos alinhamentos necessários ao modelo conceitual proposto. Outra situação de uso desse cenário ocorreu em vista da necessidade de implementação de regras e restrições por meio de axiomas para garantir a consistência do modelo diante do domínio de anotação multimídia.

Por fim, o cenário 8 – *reestruturação de recursos ontológicos* foi usado para a reorganização da conceituação resultante do mapeamento semântico envolvendo as ontologias selecionadas para reúso. A atividade de reestruturação incluiu remoção de conceitos desnecessários à conceituação, inclusão de novos conceitos e relacionamentos na estrutura e especializações adequadas que contenham em si tais elementos.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

A parte central desta seção está na apresentação do produto final da pesquisa correspondente à apresentação da especificação do modelo conceitual baseado em ontologias para o domínio de anotação de recursos multimídia em rede. Assim, a primeira subseção se incube de apresentar o resultado da análise comparativa das ontologias diante de dimensões para reúso que culminou na geração de um *ranking* de ontologias para este propósito; e a segunda subseção apresenta a especificação do Modelo Ontológico de Referência Multimídia fundamentado em recursos ontológicos desse *ranking*.

CONSIDERAÇÕES GERAIS SOBRE A ANÁLISE COMPARATIVA NO ASPECTO REÚSO

O reúso de recursos de conhecimento disponíveis para modelar o conhecimento de um domínio é uma prática recomendada pela área de Engenharia de Ontologias (SILVA; SOUZA; ALMEIDA, 2008). Um recurso ontológico abrange, a título de ilustração, ontologias já definidas ou partes de ontologias disponíveis e úteis à resolução de problemas. Orienta-se também inspecionar o conteúdo e a granularidade das ontologias, a fim de verificar o grau de cobertura dos requisitos funcionais especificados na etapa de aquisição de conhecimento.

Alguns aspectos funcionais, como nomes advindos de vocabulários bem estruturados (padrões de metadados, taxonomias, por exemplo) e modularizações de interesse, podem facilitar a extração do conhecimento requerido para reúso. Outro fator importante para a viabilização do processo de alinhamento e combinação de recursos é a linguagem de implementação da ontologia candidata a reúso, que necessita se comprometer ontologicamente com o modelo que está sendo construído para ser suficientemente expressiva na caracterização da conceituação de seu domínio. Nesse sentido, os resultados discutidos a seguir, que implicam dimensões concernentes a reúso, compreendem esse olhar analítico diante das ontologias analisadas.

O quadro 1 busca esclarecer os resultados de análise das nove ontologias candidatas a reúso ao evidenciar a estratificação da avaliação dos dezessete critérios correspondentes em cada dimensão. As pontuações inerentes a cada uma delas também são exibidas.

O gráfico 1 mostra a atuação das ontologias ante às quatro dimensões: esforço de reúso dos recursos, esforço de entendimento dos recursos, esforço de integração dos recursos e confiabilidade dos recursos.

De maneira geral, levando-se em conta as quatro dimensões, a Media Ontology, de acordo com o gráfico 1, manteve-se como a ontologia de qualidade diferenciada em relação às nove analisadas como candidatas a reúso. Acredita-se que, por ser uma proposta oriunda de um grupo de pesquisa do W3C (*Media Annotation Working Group*), especializado em questões de anotação semântica de mídias na Web, a equipe envolvida buscou empreender métricas de qualidade na construção da ontologia.

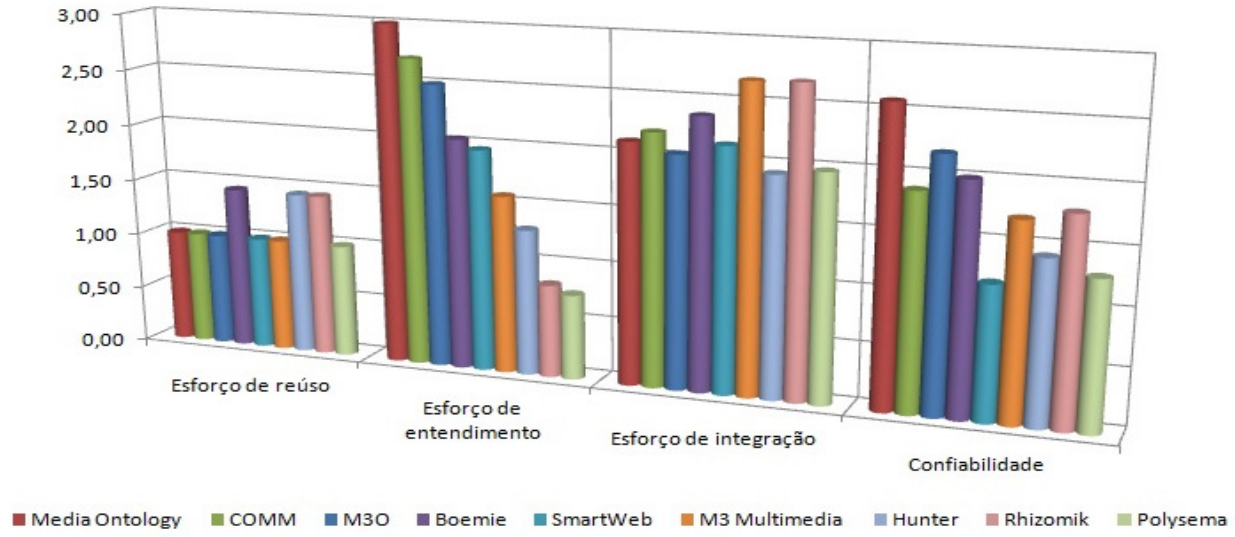
Em contrapartida, a ontologia Polysema mostrou-se a de menor qualidade em aspectos de reúso. Percebe-se que a M3 Multimedia e a Rhizomik apresentaram índices favoráveis bem próximos de esforço de integração. Já a M3O e a COMM, como propostas de modelagem mais ousadas para anotação multimídia, mantiveram-se próximas em relação às pontuações para reúso.

Quadro 1 – Resultado dos critérios avaliados nas ontologias para anotação multimídia

Critérios	Pesos	Valores																			
		Media Ontology		COMM		Boemie		M3 Multimedia		M3O		MPEG-7 Hunter		MPEG-7 Rhizomik		SmartWeb MPEG-7		Polysema MPEG-7			
Esforço de reúso dos recursos																					
Custo econômico	(-)	9	B	1	B	1	B	1	B	1	B	1	B	1	B	1	B	1	B	1	
Tempo requerido	(-)	7	B	1	B	1	M	2	B	1	B	1	M	2	M	2	B	1	B	1	
			1,00		1,00		1,44		1,00		1,00		1,44		1,44		1,00		1,00		
Esforço de entendimento dos recursos																					
Qualidade da documentação	(+)	8	A	3	A	3	A	3	B	1	A	3	B	1	B	1	M	2	B	1	
Disponibilidade de conhecimento externo	(+)	7	A	3	A	3	B	1	B	1	B	1	B	1	B	1	D	0	D	0	
Claridade no código	(+)	8	A	3	M	2	M	2	A	3	A	3	M	2	B	1	A	3	B	1	
Anotações na terminologia compatibilizada	(+)	5	A	3	A	3	M	2	B	1	A	3	B	1	D	0	A	3	B	1	
			3,00		2,71		2,04		1,57		2,50		1,29		0,82		1,96		0,75		
Esforço de integração dos recursos																					
Número de requisitos funcionais cobertos	(+)	10	47	1.18	58	1.45	42	1.05	80	2.00	21	0.53	43	1.08	118	2.95	41	1.03	29	0.73	
Adequação à extração de conhecimento	(+)	9	M	2	M	2	A	3	A	3	A	3	M	2	M	2	A	3	M	2	
Adequação à convenção de nomes	(+)	5	A	3	A	3	A	3	A	3	M	2	M	2	A	3	A	3	A	3	
Adequação à linguagem de implementação	(+)	7	A	3	A	3	A	3	A	3	A	3	A	3	A	3	M	2	A	3	
			2,12		2,21		2,37		2,68		2,04		1,93		2,69		2,14		1,98		
Confiabilidade dos recursos																					
Disponibilidade de testes	(+)	8	A	3	D	0	D	0	D	0	M	2	D	0	D	0	D	0	D	0	
Avaliação de testes	(+)	8	A	3	D	0	D	0	D	0	B	1	D	0	D	0	D	0	D	0	
Reputação do time de desenvolvimento	(+)	8	A	3	A	3	A	3	A	3	M	2	A	3	A	3	M	2	M	2	
Confiabilidade no propósito	(+)	3	A	3	A	3	A	3	A	3	A	3	M	2	M	2	A	3	A	3	
Suporte prático	(+)	7	A	3	A	3	A	3	B	1	M	2	M	2	M	2	B	1	B	1	
Recursos de conhecimento utilizados	(+)	8	M	2	A	3	A	3	A	3	A	3	A	3	A	3	A	3	A	3	
Axiomas na terminologia compatibilizada	(+)	6	B	1	M	2	A	3	A	3	A	3	D	0	A	3	D	0	B	1	
			2,58		1,88		2,00		1,71		2,19		1,42		1,79		1,17		1,29		
	Pontuação (+)		2.56		2.19		2.12		1.95		2.23		1.53		1.80		1.66		1.35		
	Pontuação (-)		1.00		1.00		1.44		1.00		1.00		1.44		1.44		1.00		1.00		
	Pontuação (=)		1.56		1.19		0.68		0.95		1.23		0.09		0.36		0.66		0.35		

Fonte: Elaborado pela autora.

Gráfico 1 – Atuação das ontologias nas dimensões para reúso



Fonte: Elaborado pela autora.

O esforço de reúso dos recursos, enquanto dimensão que influencia negativamente o *ranking*, manteve-se estável para a maioria das ontologias (conforme indica o gráfico 1). Dentro dessa dimensão, o critério custo econômico, de forma geral, foi avaliado como baixo pelo fato de o acesso às nove ontologias ter ocorrido de maneira gratuita, por meio de repositórios indicados na literatura ou de *links* apontados por máquinas de busca da Web Semântica. Já o aspecto tempo requerido variou entre baixo e médio.

As ontologias avaliadas com tempo baixo para acesso e abertura no Protégé (editor de ontologias utilizado nas análises) foram prontamente analisadas. Outras foram avaliadas com valor médio em função de alguns impasses no acesso às suas bases de conhecimento.O esforço de entendimento dos recursos é a dimensão que apresenta pontuações mais baixas para as ontologias analisadas.

Contribuíram para essa realidade a Polysema MPEG-7, a Rhizomik, a MPEG-7 Hunter e a M3 Multimedia, em geral, por motivos de escassez de fontes documentais e/ou ausência de anotações, ou mesmo de contribuição semântica, nos elementos de suas estruturas. Tal constatação as desfavorece, no aspecto consumo de tempo para se conseguir entender seus propósitos, escopos e conceituações, visando a alinhamentos consistentes.

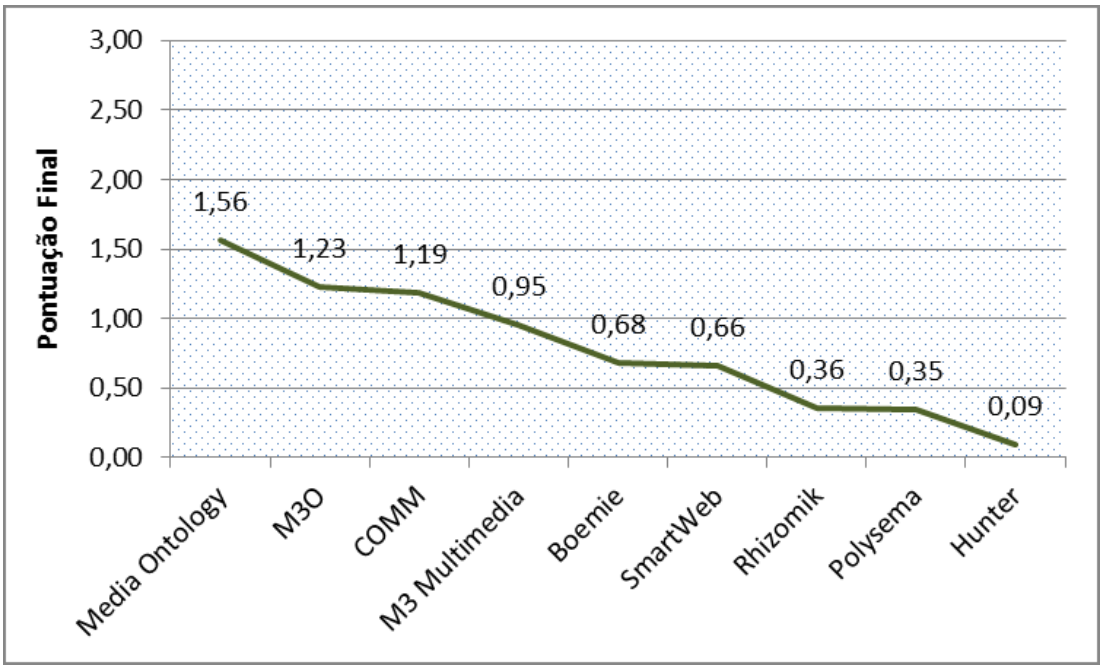
O esforço de integração dos recursos é a dimensão mais bem qualificada em comparação com as dimensões esforço de entendimento e confiabilidade. Percebe-se, no gráfico 1, que as ontologias mantiveram uma pontuação equilibrada, visto que as regras de mensuração para os critérios correspondentes (número de requisitos funcionais cobertos, adequação a extração de conhecimento e adequação a convenção de nomes) se ajustaram ao uso do padrão de metadados MPEG-7 e grande parte das ontologias analisadas tem esse padrão como referência em suas conceituações.

A OWL foi preponderante na avaliação do critério relacionado à linguagem de implementação, por sua semântica formal mapeada em lógica descritiva e com suporte a inferências lógicas, dando, portanto, condições de tratamento a questões voltadas à interoperabilidade sintática e semântica para recursos multimídia. Nesse sentido, a linguagem envolvendo a ontologia multimídia candidata a reúso seria melhor avaliada quando pertencesse à família OWL. A única ontologia que recebeu valor médio para esse critério foi a SmartWeb MPEG-7, por ser representada pela linguagem RDFS (possui limitação de expressividade em seus construtos, se comparada à OWL) em sua versão analisada. Posto isso, as avaliações correspondentes aos critérios nessa dimensão, diante das ontologias, variaram entre médio e alto. Logo, reforça-se que o método aplicado na seleção de ontologias para anotação multimídia a compor o *corpus* desta pesquisa foi bem sucedido para uma dimensão que contempla um aspecto importante relacionado à cobertura de requisitos funcionais.

A confiabilidade dos recursos pode ser considerada uma característica presente na maioria das ontologias analisadas, de acordo com o gráfico 1, em razão das seguintes constatações: *i)* todas possuem uma equipe de desenvolvimento com boa reputação*; ii)* todas são assistidas por entidades importantes no cenário mundial, tais como W3C, *European Commission*, *German Federal Ministry of Education and Research*, universidades europeias conceituadas e renomados centros de pesquisa; e *iii)* grande parte (M3O, COMM, Boemie, M3 Multimedia, Rhizomik) das ontologias se propôs a disponibilizar ricas axiomatizações em suas conceituações, as quais são fundamentadas, na maioria dos casos, em ontologias de alto nível, em padrões de projeto multimídia, e no padrão de metadados MPEG-7.

Finalmente, o gráfico 2 apresenta o *ranking* das ontologias candidatas a reúso, cujo resultado da pontuação final (ver quadro 1) foi obtido por cálculo de média ponderada, contendo em si os critérios com influência positiva e negativa no *ranking*, conforme se explanou na seção Metodologia.

Gráfico 2 − *Ranking* das ontologias candidatas a reúso



Fonte: Elaborado pela autora.

ESPECIFICAÇÃO DO MODELO ONTOLÓGICO DE REFERÊNCIA MULTIMÍDIA

A especificação de requisitos foi o ponto de partida para a proposta do modelo conceitual envolvendo ontologias para anotação multimídia. *A priori*, as definições necessárias para a especificação de requisitos seriam o propósito do Modelo Ontológico de Referência Multimídia e o seu escopo. O propósito do modelo abrange a intenção de seu uso, os possíveis cenários que demandam o uso e os usuários que irão utilizá-lo. O escopo vai incluir um conjunto de requisitos funcionais previamente determinados.

O propósito do Modelo Ontológico de Referência Multimídia é representar uma conceituação consensual e compartilhada por uma determinada comunidade para organização semântica de anotações de recursos multimídia em rede. Logo, o modelo tem a intenção de promover enriquecimento de variados tipos de metadados multimídia por meio de uma estrutura informacional de alto nível ontológico que possa servir, por exemplo, como solução de agregação de *datasets* multimídia em LOD, viabilizando interoperabilidade a nível sintático e semântico entre diversas instituições e seus sistemas de informação. Vale ressaltar que o sentido de consensual e compartilhado é o que caracteriza o modelo como um objeto de referência subjacente a esforços de pesquisas voltados a modelos e tecnologias para processamento de metadados multimídia.

O modelo conceitual poderia ser usado em sistemas de informação voltados a instituições de cultura, como arquivos, bibliotecas, museus, centros de documentação e projetos de memória, cujos usuários consomem, interpretam, manipulam e geram conteúdos multimídia nos acervos que, atualmente, se encontram digitalizados e acessíveis em portais ou repositórios na Web. Catalogadores é outra categoria de usuários que exerce um papel importante na associação de anotações em recursos multimídia, principalmente em espaços de conhecimento dinâmico que incluem os sistemas de bibliotecas digitais.

Por fim, mas sem esgotar as possibilidades de uso, portais de notícias das mais variadas naturezas necessitam de métodos eficientes para organizar conteúdos multimídia e transmiti-los de maneira inteligente a várias tipologias de usuários.

O escopo do modelo foi especificado por meio de requisitos funcionais (RF) e não funcionais (RNF) importantes à proposição de uma arquitetura abrangente para a modelagem de metadados destinados à representação de recursos multimídia a qualquer domínio de conhecimento. Os requisitos elencados são comentados a seguir:

a) (RF) Considerar o conteúdo e a realização da mídia em várias modalidades, tais como áudio, imagem, texto, modelo 3D e vídeo: a separação entre objetos de informação e suas realizações é importante, pois alguns metadados independentes de conteúdo, como tamanho do arquivo ou localização da mídia na Web, são comumente aplicados à realização da informação; já os metadados descritivos de conteúdo multimídia buscam descrever a mensagem a ser transmitida para o consumidor do conteúdo. Portanto, esta separação torna-se relevante no sentido de fornecer uma distinção clara entre a semântica do conteúdo e o recurso de mídia.

b) (RF) Cobrir metadados independentes, dependentes e descritivos de conteúdo: esquemas voltados à descrição multimídia devem contemplar aspectos de gerenciamento e administração de recursos digitais; aspectos primitivos para metadados visuais e de áudio; aspectos relacionados ao conteúdo da mídia, incluindo decomposição de tipos de mídia e suas localizações, associação de conteúdo da mídia a entidades semânticas veiculadas (oriundas de ontologias de domínio), como, a título de ilustração, um rosto de uma pessoa retratado em uma imagem; além de aspectos vinculados à personalização de conteúdo para facilitar navegação, acesso e interação de usuários em relação ao consumo de conteúdo.

c) (RNF) Possuir uma ontologia de alto nível como referência: ontologias de alto nível têm sido denominadas "Ontologias de Fundamentação" (do inglês, *Foundational Ontologies*) (GUIZZARDI, 2005), as quais descrevem conceitos bastante gerais, tais como, espaço, tempo, matéria, objeto, evento, agente, etc., e são consideradas sistemas de categorias filosoficamente bem empregadas e independentes de domínio. O seu emprego beneficia semanticamente a estrutura taxonômica central da ontologia quando esclarece o significado pretendido sobre os termos, apoiando, por exemplo, a integração de instâncias do conteúdo da mídia com ontologias de domínios específicos.

d) (RNF) Ser fundamentado em padrões multimídia estendidos de padrões de projeto de ontologias (GANGEMI; PRESUTTI, 2009): ameniza os desafios impostos na atividade de reúso com visualizações diagramáticas aceitáveis e memorizáveis para um conjunto específico de questões de competência (problema e sua solução). Algumas ontologias para anotação multimídia utilizam padrões de projeto para organização genérica de entidades e relacionamentos subjacentes ao domínio multimídia, como anotação e decomposição.

e) (RNF) Considerar ontologias bem colocadas em um *ranking* produzido a partir de critérios bem fundamentados: uso de uma metodologia madura, robusta e eficiente para análise e avaliação criteriosa de ontologias para anotação multimídia.

f) (RNF) Assegurar interoperabilidade em relação a conteúdo multimídia na Web: garante que o significado intencionado da semântica capturada possa ser compartilhado entre diferentes aplicações no escopo da Web Semântica. Além do compartilhamento semântico de conteúdo multimídia, o modelo deve prever meios de transmissão em alguma sintaxe acordada por uma comunidade, que, nesse caso, seria por meio de linguagens da Web Semântica como RDF/OWL, a título de ilustração.

g) (RNF) Possuir uma arquitetura extensível em relação à construção de uma ontologia multimídia abrangente: considerando que uma ontologia está sempre em evolução, a inclusão de novos conceitos deve ser sempre prevista na conceituação com característica extensível. A extensibilidade é assegurada na medida em que padrões de projeto e ontologias de alto nível conseguem, através das metacategorias, ampliar a possibilidade de inserção de novos conceitos sem precisar modificar o modelo central subjacente.

A partir do *ranking* e das constatações alcançadas na análise comparativa, tornou-se possível selecionar e justificar os recursos ontológicos apropriados a reúso com base nos requisitos delineados anteriormente. As ontologias que melhor se ajustaram à proposição do Modelo Ontológico de Referência Multimídia foram a *Media Ontology (1,56)*, a *M3O (1,23)*, a *COMM (1,19)* e a *M3 Multimedia (0,95).*

Dado que o Modelo Ontológico de Referência Multimídia deve ser fundamentado em uma ontologia de alto nível, bem como em padrões de projeto multimídia, e, ainda, tratar diferenças semânticas entre o conteúdo e a realização da mídia, a M3O foi eleita como a ontologia que mais se adéqua a esses requisitos. A escolha se justifica pelo fato de a arquitetura da conceituação da M3O ser fundamentada na ontologia de alto nível DOLCE+DnS Ultralight (DUL) e em três padrões de projeto referenciados por esta, a saber: *Description and Situation* (DnS), *Information and Realization Pattern* e *Data Value Pattern*.

Os padrões multimídia da M3O são estendidos do padrão *Description and Situation*, contemplando *AnnotationPattern* (padrão de anotação), *DecompositionPattern* (padrão de decomposição) e *CollectionPattern* (padrão de coleção). Os seus diagramas de projeto são facilmente memorizáveis pela simplicidade de seus esquemas com poucas classes e relações, o que viabiliza o entendimento do raciocínio de modelagem empregado nas conceituações.

Além disso, os três padrões multimídia (anotação, decomposição e coleção) atuam sob a semântica especificada no padrão *Information and Realization* que representa a distinção entre objetos de informação e realizações de informação.

A Media Ontology é a ontologia recomendada para os *metadados independentes de conteúdo,* pelo fato dessa ontologia possuir um índice de cobertura satisfatório para essa categoria de metadados (em relação a outras ontologias analisadas), com destaque para descritores alinhados com o padrão Dublin Core.

As ontologias COMM e M3 Multimedia oferecem descritores que se alinham bem aos *metadados dependentes de conteúdo*, úteis quando do processamento computacional de dados digitais para geração automática de metadados. Ambas apresentaram índices bem próximos para cobertura visual, em destaque para descritores envolvendo cor, textura, forma e localização de regiões de interesse. Metadados para descrever características 3D, por exemplo, podem ser encontrados nas duas ontologias, especialmente nos aspectos visuais relacionados à forma, visto que ambas se fundamentam no padrão MPEG-7 para descrição de conteúdo multimídia. O padrão MPEG-7 contempla descritores para características tridimensionais de objetos, como simetria, circularidade, localização de eixos, tamanho e orientação de segmentos consecutivos de bordas, pontos de curvaturas e ângulos de curvas. Os recursos de conhecimento relacionados aos metadados de áudio podem ser selecionados da M3 Multimedia pelo fato desta ter praticado reúso tanto dos metadados visuais quanto dos de áudio da ontologia VDO Boemie.

Os *metadados descritivos de conteúdo* voltados à semântica são, geralmente, ligados a instâncias de ontologias de domínio cujos rótulos semânticos são organizados na estrutura taxonômica de uma ontologia de fundamentação.

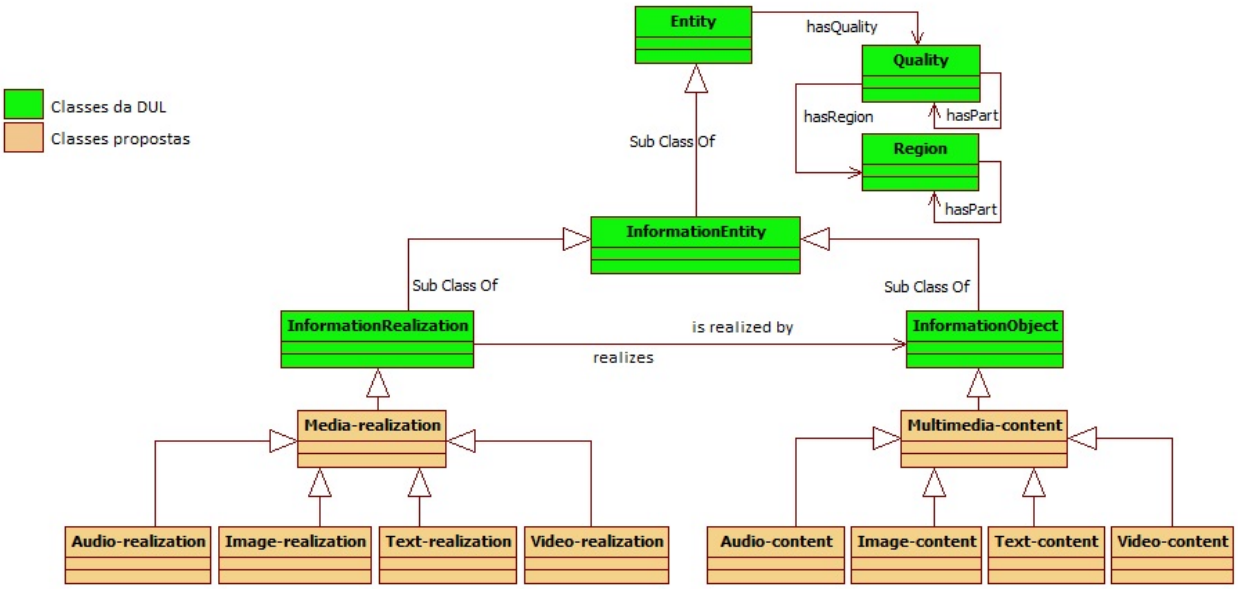
Como a M3O se integra à ontologia de fundamentação DUL, esta consegue cumprir o papel de organizar rótulos semânticos advindos de ontologias de domínios específicos em entidades como evento, objeto, tempo, lugar, etc, e, ainda, tratar de seus relacionamentos. A M3 Multimedia cobre propriedades para aspectos de navegação e acesso (personalização de conteúdo), descritores de áudio com características de alto nível (conteúdo falado, por exemplo), e descritores comuns para anotação de segmentos.

Torna-se relevante assinalar também que a M3O, por possuir uma ontologia de alto nível como referência e, ainda, ser fundamentada em padrões multimídia estendidos de padrões de projeto de ontologias, possui todas as características que asseguram o atendimento aos requisitos não funcionais delineados, como tratamento de diferentes níveis de granularidade, interoperabilidade, separação de interesses e extensibilidade. Uma vez realizados os alinhamentos pré-estabelecidos com os requisitos, foi possível propor a organização semântica dos recursos de conhecimento trabalhados na pesquisa em arranjos ou agrupamentos, incluindo: i) tipos de mídia e suas realizações envolvidas no contexto multimídia; e ii) categorias específicas de metadados multimídia (advindas especialmente do padrão MPEG-7) e suas respectivas classes ontológicas, organizadas por categoria de metadados independentes, dependentes e descritivos de conteúdo.

No primeiro arranjo (figura 1), os tipos de mídia e a suas realizações foram agrupados nas classes conceituais nomeadas, respectivamente, *Multimedia-content* e *Media-realization*. Tais classes foram então generalizadas para as classes da DUL (ontologia de fundamentação da M3O) correspondentes a *Information object* e a *Information realization*, respectivamente. Novas especializações de mídias podem ser concebidas para aplicações específicas, dada a natureza extensiva do modelo.

Os quadros 2 e 3 apresentam os demais arranjos propostos para organizar os recursos de conhecimento diante dos tipos de metadados multimídia trabalhados na pesquisa.

Figura 1 – Classes de entidades centrais do Modelo Ontológico de Referência Multimídia



Fonte: Elaborado pela autora.

Quadro 2 – Arranjos para os tipos de metadados independentes e dependentes de conteúdo

Metadados Independentes de Conteúdo		
Ontologia Media Ontology		
Categoria de Metadados	Classe na ontologia	Descrição da classe
Criação e produção da mídia	Media_Creation (*)	Descreve características envolvendo a criação do conteúdo da mídia e de recursos a ele associados.
Classificação da mídia	Media_Classification (*)	Descreve características destinadas à classificação da mídia, tais como gênero, assunto, propósito, idioma, dentre outras.
Informação da mídia	Media_Information (*)	Descreve os meios de armazenamento contemplando formato, compressão e codificação do conteúdo.
Uso da mídia	Media_Usage (*)	Descreve características que refletem direitos de uso, registro e disponibilidade de uso da mídia.
Metadados Dependentes de Conteúdo		
Ontologia COMM		
Categoria de Metadados	Classe na ontologia	Descrição da classe
Visuais	structured-data-parameter. visual-descriptor-parameter	Descreve características visuais primitivas para cor, textura, forma e movimento.
Cor	.color- descriptor-parameter	Descreve vários descritores e parâmetros de apoio na representação de diferentes aspectos de características envolvendo cor.
Textura	.texture-descriptor-parameter	Descreve aspectos importantes na revelação de características tácteis, de profundidade e orientação de superfícies para uma imagem.
Forma	.shape-descriptor-parameter	Descreve características relacionadas a arranjo espacial de pontos (pixels) que pertencem a um objeto ou uma região. Os descritores podem ser agrupados em classes 2D ou 3D.
Movimento	.motion-descriptor-parameter	Descreve características espaciais e temporais capturadas pelo movimento de câmera, objeto em movimento, ou ambos.
Localização	localization-descriptor-parameter	Descreve localização para regiões de interesse em domínios espacial e espaço temporal.

(Continua)

Quadro 2 – Arranjos para os tipos de metadados independentes e dependentes de conteúdo (Conclusão)

Ontologia M3 Multimedia		
Categoria de Metadados	Classe na ontologia	Descrição da classe
Áudio	LL_Audio_Descriptor	Descreve descritores primitivos envolvendo características espectrais, paramétricas e temporais para descrever sinais e arquivos de áudio.
Base Espectral	.Spectral_Basis_Descriptor	Descreve projeções de baixa dimensionalidade de um espaço espectral de alta dimensão para ajudar na compacidade e identificação.
Timbre Espectral	.Spectral_Timbral_ Descriptor	Descreve características do timbre relacionadas ao espectro do sinal.
Timbre Temporal	.Temporal_Timbral_ Descriptor	Descreve características temporais de segmentos de áudio; úteis especialmente para descrever características do timbre de instrumentos musicais.
Paramétricos de Sinal	.Signal_Parameter_ Descriptor	Descreve sinais periódicos ou quase periódicos.
Espectral Básico	.Basic_Spectral_Descriptor	Descreve descritores derivados da análise de frequência do sinal.
Básico	.Basic_Descriptor	Descreve descritores básicos de uso geral e aplicáveis a todos os tipos de sinais.

Fonte: Elaborado pela autora.

Quadro 3 – Arranjos para os tipos de metadados descritivos de conteúdo

Metadados Descritivos de Conteúdo		
Ontologia M3 Multimedia		
Categoria de Metadados	Classe na ontologia	Descrição da classe
Elementos comuns para segmentos	Segment_Label (*)	Descreve elementos comuns para anotação de segmentos, tais como aspectos semânticos abstraídos do conteúdo, local de acesso, criador, informação da mídia e licença de uso.
Navegação e acesso	Navigation_Access (*)	Descreve aspectos de características que facilitam a navegação e o acesso a conteúdo multimídia, a exemplo dos sumários.
Áudio de alto nível	HL_Audio_Descriptor	Descreve canonicamente um som com certo grau de generalidade, incluindo descritores voltados à cobertura de domínios específicos.
Conteúdo falado	.Spoken_Content_ Descriptor	Descreve detalhes das palavras faladas dentro de um fluxo de áudio.
Ontologia M3O		
Categoria de Metadados	Classe na ontologia	Descrição da classe
Organização de objetos digitais em coleções	CollectionPattern	Descreve características de coleções de entidades de informação com propriedades comuns.
Segmentos de mídia	DecompositionPattern	Descreve a estrutura de conteúdo multimídia em termos de segmentos, tais como quadros, regiões em movimento, regiões estáticas e faixas de áudio.
Semântica de conteúdo	DUL:Entity	Descreve objetos, eventos e noções do mundo real que podem ser abstraídos do conteúdo multimídia.
Padrão de metadados MPEG-7		
Categoria de Metadados	Classe na ontologia	Descrição da classe
Segmento temporal	Temporal_Segment (*)	Descreve um conjunto de características temporais relacionadas à decomposição de segmentos para conteúdo de mídias específicas, tais como vídeo, áudio, cena, audiovisual e região em movimento.
Segmento espacial	Spatial_Segment (*)	Descreve um conjunto de características espaciais relacionadas à decomposição de segmentos para conteúdo de mídias específicas, tais como imagem 2D, imagem 3D e região em movimento.
Segmento espaço-temporal	Spatio_Temporal_ Segment (*)	Descreve um conjunto de características espaço temporal relacionadas à decomposição de segmentos para conteúdo de mídias específicas, tais como região em movimento e região audiovisual.

Fonte: Elaborado pela autora.

Para cada ontologia envolvida no reúso, bem como para alguns conceitos advindos do padrão MPEG-7, foram mapeadas (ou propostas) classes conceituais correspondentes. Salienta-se que a nomenclatura das classes foi mantida conforme a sua origem ontológica. Já as classes propostas foram nomeadas seguindo convenções terminológicas subjacentes aos casos de uso pesquisados. Por fim, o símbolo (*) indica uma nova classe para o modelo.

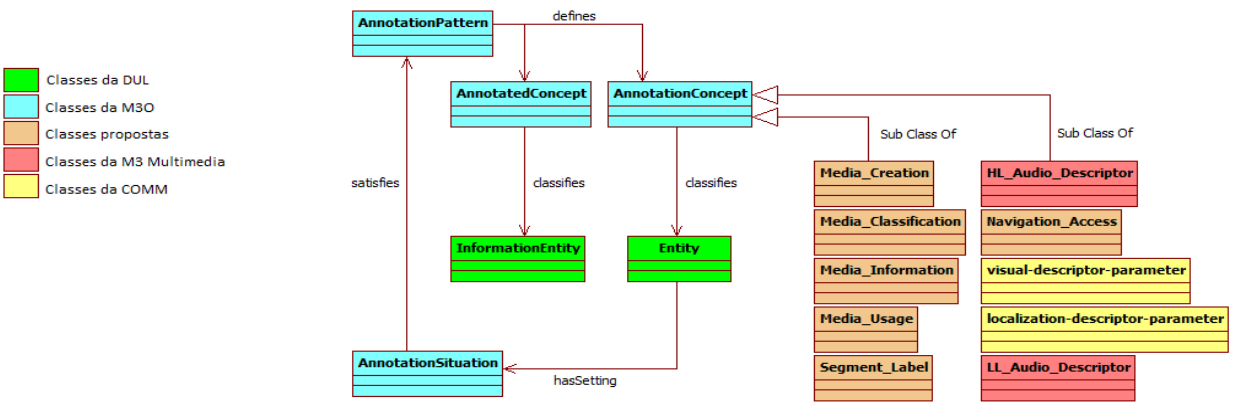
As ontologias que apresentaram modelagens baseadas em padrões (como a COMM e a M3O) já possuíam agrupamentos formais e coerentes (proporcionados pelos axiomas) representados em classes conceituais. Já as ontologias que apresentaram modelagens centradas em relações e atributos (como a Media Ontology) tiveram suas propriedades agrupadas em classes conceituais. Para tais casos, tornou-se necessária a criação de axiomas para o tratamento dos agrupamentos, no sentido de se declarar formalmente os descritores pertencentes a cada categoria de metadados, como, por exemplo, a classe "Media Creation", proposta no modelo (indicada no quadro 2), que deveria possuir algum título, um local de acesso ou uma data. Após a proposição dos arranjos envolvendo os recursos de conhecimento da pesquisa, tornou-se possível efetuar o mapeamento semântico das classes ontológicas indicadas nos quadros 2 e 3 para os padrões de projeto multimídia da M3O (*AnnotationPattern*, *DecompositionPattern* e *CollectionPattern*), conforme se apresenta nos diagramas de classes elucidados a seguir.

Vale reassaltar que, para fins de melhor visualização e entendimento, o Modelo Ontológico de Referência Multimídia foi segmentado em três partes associadas aos padrões de projeto multimídia subjacentes à conceituação proposta. Os diagramas apresentam classes de âmbito mais genérico e não expõem, portanto, classes específicas, por razões de simplificação na forma de visualização dos mesmos.

Na M3O, um *AnnotatedConcept* classifica uma *InformationEntity*, que é o recurso de informação a ser anotado ou o sujeito da anotação, e cada item de metadados é representado por uma *Entity*, a qual é classificada por um *AnnotationConcept*. Os mapeamentos se sucederam a partir dessa lógica conceitual, que descreve uma entidade de informação (que pode ser um objeto e uma realização da informação) e os metadados que participam do processo de anotação. As classes (oriundas de agrupamentos ou mapeadas das ontologias correspondentes) referentes aos metadados multimídia foram especializadas na classe *AnnotationConcept*, que atribui às entidades de dados o papel de anotação e que especifica a sua interpretação como metadados.

O diagrama exibido na figura 2 apresenta o mapeamento semântico para o padrão de anotação ora descrito.

Figura 2 – Classes de entidades de anotação do Modelo Ontológico de Referência Multimídia

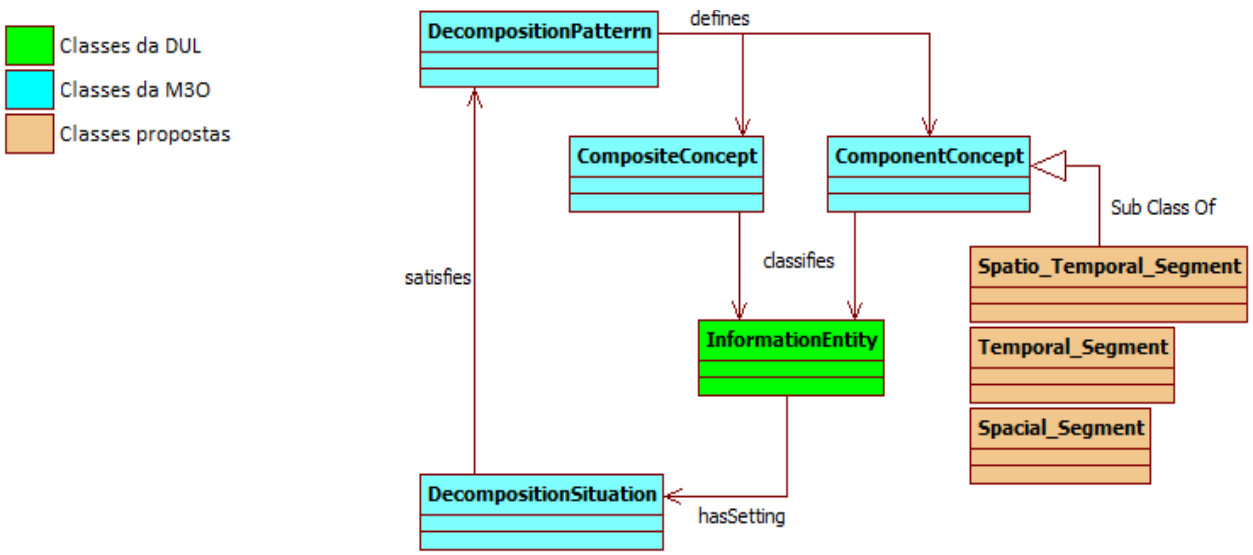


Fonte: Elaborado pela autora.

Os mapeamentos que abrangem o padrão *DecompositionPattern* são assegurados pelas classes de tipos de mídia alinhadas com as entidades de informação da DUL, conforme se apresentou na figura 1. Assim, um *CompositeConcept* faz o papel da mídia envolvida na decomposição e o *ComponentConcept* faz o papel dos segmentos de mídia resultantes da decomposição. No caso de *CompositeConcept*, o emprego de declarações do tipo *owl:disjointWith* para restringir a participação de instâncias de tipos de mídia em classes indevidas torna-se crucial, como ilustra um *Audio-content*, que é disjunto de *Video-content* e de *Image-content*. Já para *ComponentConcept*, as classes propostas para os tipos de segmentos resultantes de um processo de decomposição foram especializadas como subclasses de *ComponentConcept* nomeadas, *TemporalSegment*, *SpatialSegment* e *SpatioTemporalSegment*, representando, respectivamente, características de dimensões temporais, espaciais e espaço temporais desses segmentos. Para tais classes, torna-se relevante a modelagem de axiomas para impor restrições sobre os tipos de segmentos que formam decomposições válidas para conteúdos que contém uma mídia em específico. Exemplo disso, seria um segmento correspondente a uma região em movimento (especializado da classe de componente *SpatioTemporalSegment*) ser classificado somente no tipo de mídia vídeo; além de suas localizações serem endereçadas apenas às classes de anotação concernentes a metadados de localização visual e de tempo. Observa-se, desse modo, a participação do padrão de anotação nas descrições dos segmentos resultantes. A classe proposta para classificar entidades com papéis de metadados comuns para segmentos foi nomeada *Segment_Label*, conforme mostrado na figura 2, que pode comportar, inclusive, rótulos semânticos (ou instâncias) advindos de ontologias de domínio numa situação de anotação de conteúdo semântico. Para tais entidades, a ontologia de fundamentação DUL se incumbe de organizar as suas naturezas semânticas por meio das classes de alto nível, como objeto, evento, agente, tempo e lugar.

O diagrama de classes da figura 3 apresenta o mapeamento semântico para o padrão de decomposição ora descrito.

Figura 3 – Classes de entidades de decomposição do Modelo Ontológico de Referência Multimídia



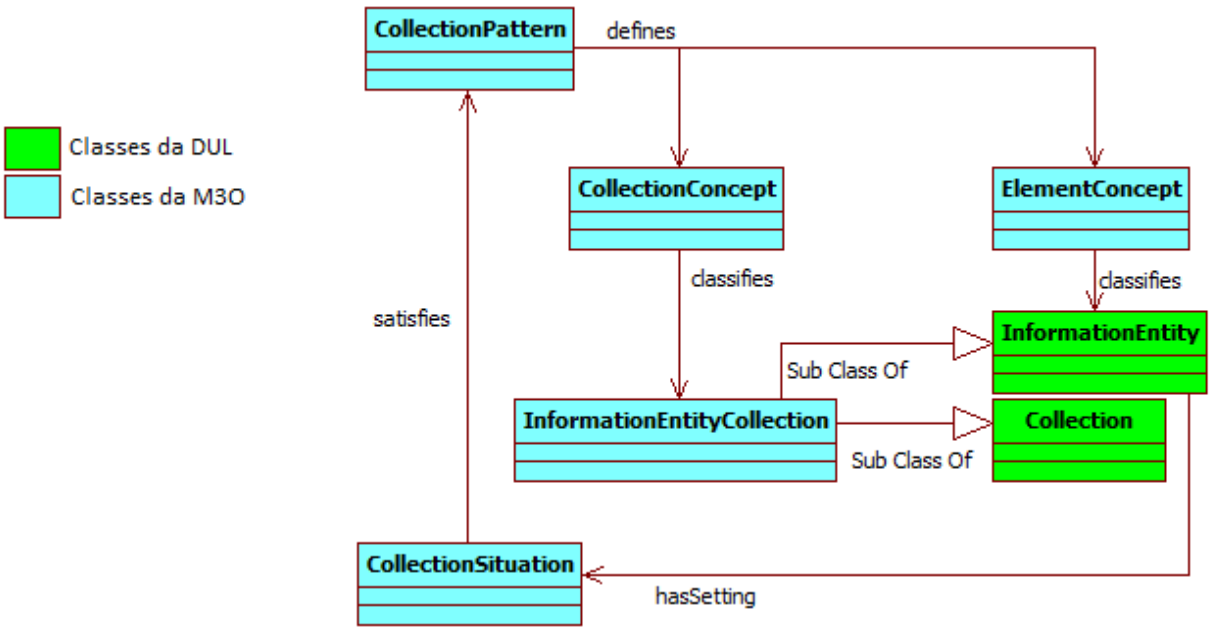
Fonte: Elaborado pela autora.

Finalmente, o padrão multimídia para coleção da M3O permite representar coleções de entidades de informação com propriedades comuns por meio do *CollectionPattern*, que apoia a criação colaborativa de coleções, levando em consideração a fonte ou origem das entidades de informação envolvidas. Os principais conceitos do padrão de coleção estabelecem especializações com os padrões de projeto da DUL.

A classe *CollectionPattern* define que existe exatamente um *CollectionConcept*, que classifica uma *InformationEntityCollection*, que, por sua vez, é uma coleção de entidades de informação. O *AnnotationPattern* interage no sentido de prover classes de metadados multimídia às *InformationEntityCollection*.

O diagrama de classes da figura 4 apresenta o mapeamento semântico para o padrão de coleção ora descrito.

Figura 4 – Classes de entidades de coleção do Modelo Ontológico de Referência Multimídia



Fonte: Elaborado pela autora.

Uma vez que os mapeamentos semânticos tenham sido finalizados, recomenda-se as seguintes tarefas baseadas em metodologias de engenharia de ontologias operacionalizadas no decorrer da pesquisa: i) remoção de conceitos desnecessários na conceituação da ontologia resultante, a fim de se evitar uma taxonomia extensa e com conceitos ambíguos; ii) documentação clara e precisa de todos os elementos ontológicos constituintes; e iii) validação da taxonomia para verificar a consistência da ontologia resultante.

Para assegurar que a modelagem do domínio possa ser realizada de maneira satisfatória, recomenda-se a participação de especialistas que lidam com arquivos multimídia. Profissionais de Biblioteconomia e Ciência da Informação, por exemplo, são especialistas em catalogação descritiva e indexação de recursos de informação e podem contribuir para o estabelecimento de acordos sobre metadados específicos para cada entidade de informação envolvida na especificação do modelo.

Profissionais especialistas em visão computacional, processamento de imagens e sinais de áudio podem contribuir nas decisões de modelagem concernentes a metadados depedentes de conteúdo, os quais são altamente subordinados ao conhecimento técnico de tais áreas. Desse modo, o ontologista pode se ocupar, principalmente, de tarefas relacionadas à organização semântica do sistema de informação.

CONSIDERAÇÕES FINAIS

O Modelo Ontológico de Referência Multimídia traz inovações e avanços ao campo da Ciência da Informação, especialmente à área de Catalogação Descritiva em ambientes digitais juntamente com o emprego de vocabulários semânticos na representação consistente de conteúdos associados a objetos digitais ou multimídia em rede. Nesse sentido, evidencia-se a evolução do conceito "metadados" para anotações semânticas cujo arcabouço teórico e conceitual encontra-se nos princípios da Web Semântica e seu conceito mais recente de *Linked Open Data*.

O modelo proposto foi construído a partir do uso do guia *Neon Methodology* que se mostrou robusto e eficiente na explicitação das diferentes dimensões e variabilidades na análise dos recursos de conhecimento identificados na literatura e em repositórios da Web Semântica. A análise e comparabilidade de vocabulários semânticos promoveram as condições necessárias para a seleção e o reúso de recursos de conhecimento apropriados a representar uma estrutura abrangente capaz de modelar variados tipos de metadados (independentes, dependentes e descritivos de conteúdo) no intuito de organizá-los semanticamente para fins de anotação de recursos multimídia. Assim sendo, características relevantes que podem e devem ser descritas para melhor recuperação de recursos multimídia advindas de padrões ISO, como o MPEG-7 e o Dublin Core, agregadas a uma camada semântica dos dados e metadados são benefícios, em tese, almejados pelo uso do Modelo Ontológico de Referência Multimídia na organização da informação em ambientes digitais diversos.

Somado a isso, o modelo enseja descrições padronizadas e livres de ambiguidade para os itens constantes nas bases de dados institucionais, o que as torna interoperáveis e, consequentemente, mais adequadas a procedimentos de integração de dados heterogêneos. Nesse sentido, o modelo pode viabilizar a entrada dessas instituições na rede LOD com o propósito de publicização de seus recursos digitais ou multimídia expressos em conceitos e relacionamentos ontológicos de alto nível, permitindo, portanto, o compartilhamento e a ligação entre vários *datasets* multimídia na Web.

Com relação às três questões que se delinearam na presente pesquisa, o Modelo Ontológico de Referência Multimídia conseguiu respondê-las quando empregou, em sua estrutura central, uma ontologia conceitualista baseada em aspectos cognitivos, filosóficos e linguísticos, que fornecem metacategorias para descrição formal de eventos, objetos, tempo, espaço, dentre outras, responsáveis por organizar semanticamente conteúdos advindos de ontologias de domínios específicos. Nesse viés, a semântica formal oriunda da linguagem de representação OWL contribui sobremaneira para o nível de abrangência da estrutura conceitual proposta, em que se busca descrever qualquer aspecto relacionado a dado multimídia. A abrangência é alcançada pela utilização de princípios da engenharia de ontologias, a qual sugere o emprego de ontologias de fundamentação e padrões de projeto de conteúdo ontológico. Desse modo, o Modelo Ontológico de Referência Multimídia busca viabilizar sua ligação com ontologias de domínios específicos por meio de definições axiomatizadas de conceitos de alto nível, oriundos da ontologia de fundamentação *DOLCE+DnS Ultralight*, e seus padrões de projeto *Description and Situation*, *Information and Realization* e *Data Value*, usados para a organização genérica de entidades associadas a conteúdo multimídia, como anotação, decomposição e coleção. Assim sendo, a estrutura taxonômica central do modelo proposto propicia a solução da primeira questão de pesquisa, que é saber *como expressar a estrutura conceitual subjacente à descrição da realidade documental de tipo multimídia*.

A segunda questão, que corresponde a saber *qual a estratégia para selecionar e alinhar eficientemente vocabulários e ontologias multimídia concebidos por comunidades distintas, de modo a cobrir satisfatoriamente aspectos ontológicos da realidade documental de tipo multimídia*, é solucionada pelo *ranking* obtido a partir de uma avaliação criteriosa de dimensões concernentes a reúso de ontologias, o que assegurou a seleção de recursos de conhecimento adequados para a integração na proposta de conceituação do Modelo Ontológico de Referência Multimídia.

A terceira questão da pesquisa, investigar *como organizar sistematicamente tipologias de metadados existentes para descrição de conteúdo multimídia diante de variados contextos e necessidades*, foi solucionada pela cobertura de requisitos funcionais proporcionada pelo modelo ontológico, embasada em categorias de metadados (independentes, dependentes e descritivos de conteúdo) que, por sua vez, buscam ampliar e enriquecer os pontos de acesso para melhorar a gestão, a organização e a recuperação de objetos digitais em variados contextos e conjunturas.

Como desdobramentos de trabalhos futuros, pretende-se validar o Modelo Ontológico de Referência Multimídia proposto em estudos de integração de acervos multimídia disponíveis na rede LOD e, a partir dessa integração, implementar e testar várias consultas com mecanismos de inferências úteis, a fim de se obter resultados mais conclusivos. Por fim, esse trabalho inaugura uma profícua linha de pesquisa, a qual se pretende dedicar nos próximos anos.

REFERÊNCIAS

ADJEROH, D. A.; NWOSU, K. C. Multimedia database management: requirements and issues. *IEEE Multimedia*, [*s.l.*], v. 4, n. 3, p. 24-33, July/Sept 1997. DOI: https://doi.org/10.1109/93.621580.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. *Scientific American*, [*s.l.*], v. 284, n. 5, p. 34-43, May. 2001. Disponível em: https://www.jstor.org/stable/26059207. Acesso em: mar. 2021.

GANGEMI, A.; PRESUTTI, V. Ontology design patterns. *In:* STAAB, S.; STUDER, R. (ed.). *Handbook on ontologies*. Berlin: Springer, 2009. p. 221-243.

GUIZZARDI, G. *Ontological foundations for structural conceptual models*. 2005. Thesis (Telematics and Information Technology PhD) - Universidade de Twente, Enschede, Holanda, 2005. Disponível em: https://research.utwente.nl/en/publications/ontological-foundations-for-structural-conceptual-models. Acesso em: mar. 2021.

HYVÖNEN, E. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. [*S.l.*]: Morgan&Claypool, 2012. (Synthesis Lectures on the Semantic Web: Theory and Technology, v. 2).

LEMOS, D. L. S.; MENDONÇA, F. M.; SOUZA, R. R. Ontologias no suporte semântico na organização de acervos digitais em rede. *In:* ALMEIDA, M. B. (org.). *Representação do Conhecimento, Ontologias e Linguagem*: pesquisa aplicada em Ciência da Informação. Curitiba: CRV, 2020. p. 161-191.

LEMOS, D. L. S.; SOUZA, R. R. Ontologias na representação de documentos: um panorama atual para descrição de conteúdo multimídia em rede. *Inf. & Soc.*: Est., João Pessoa, v. 29, n. 4, p. 103-134, out./dez. 2019. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/47421. Acesso em: mar. 2021.

LEMOS, D. L. S.; SOUZA, R. R. Representação de recursos multimídia na web: uso e reúso de padrões de anotação. *Perspectivas em Ciência da Informação,* Belo Horizonte, v. 25, n. especial, p. 202-232, 2020. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/4305. Acesso em: mar. 2021.

MACHADO, L. M. O.; SOUZA, R. R.; SIMÕES, M. G. Semantic web or web of data? a diachronic study (1999 to 2017) of the publications of tim berners-lee and the world wide web consortium. *JASIST*, [*s.l.*], v. 70, n. 7, p. 701-714, 2019. DOI: https://doi.org/10.1002/asi.24111. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24111. Acesso em: mar. 2021.

POTENZIANI, M. *et al*. Publishing and Consuming 3D Content on the Web: a Survey. *Foundations and Trends in Computer Graphics and Vision*, [*s.l.*], v. 10, n.4, p. 244-333, 2018. DOI: http://dx.doi.org/10.1561/0600000083. Disponível em: http://vcg.isti.cnr.it/Publications/2018/PCDS18. Acesso em: 15 abr. 2020.

SALEMBIER, P.; SMITH, J. MPEG-7 multimedia description scheme. *IEEE Transactions on Circuits and Systems for Video Technology*, [*s.l.*], v. 11, n. 6, Jun. 2001. DOI: 10.1109/76.927435. Disponível em: https://ieeexplore.ieee.org/document/927435. Acesso em: mar. 2021.

SCHANDL, B. *et al*. Linked Data and multimedia: the state of affairs. *Multimedia Tools and Applications*, [*s.l.*], v. 59, p. 523-556, 2012. DOI 10.1007/s11042-011-0762-9. Disponível em: https://link.springer.com/article/10.1007/s11042-011-0762-9. Acesso em: mar. 2021.

SHADBOLT, N., BERNERS-LEE, T.; HALL, W. The semantic web revisited. *IEEE Intelligent Systems*, [*s.I.*], v. 21, n. 3, p. 96-101, 2006. DOI 10.1109/MIS.2006.62. Disponível em: https://ieeexplore.ieee.org/document/1637364. Acesso em: mar. 2021.

SILVA, D. L. *Ontologias para representação de documentos multimídia*: análise e modelagem. 2014. 441 f. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2014. Disponível em: http://hdl.handle.net/1843/BUOS-9NCGYM. Acesso em: mar. 2021.

SILVA, D. L; SOUZA, R. R.; ALMEIDA, M. B. Ontologias e vocabulários controlados: comparação de metodologias para construção. *Ci. Inf.*, Brasília, v. 37, n. 3, p. 60-75, set./dez. 2008. Disponível em: http://revista.ibict.br/ciinf/article/view/1204. Acesso em: mar. 2021.

SILVA, D. L.; SOUZA, R. R. Representação de documentos multimídia: dos metadados às anotações semânticas. *Tendências da Pesquisa Brasileira em CiênciaS da Informação*, [*s.l.*], v. 7, n. 1, 2014. Disponível em: https://brapci.inf.br/index.php/res/v/119482. Acesso em: mar. 2021.

SUÁREZ-FIGUEROA, M. C.; GÓMEZ-PÉREZ, A.; FERNÁNDEZ-LÓPEZ, M. The NeOn methodology for ontology engineering. *In:* SUÁREZ-FIGUEROA, M. C. *et al.* (ed.). *Ontology Engineering in a Networked World*. Berlin: Springer, 2012. p. 9-34.

Aplicação de Dados Governamentais Abertos à luz da ciência da informação

Marckson Roberto Ferreira de Sousa
Doutorado em Engenharia Elétrica pela Universidade Federal da Paraíba (UFPB) – PB - Brasil. Professor da Universidade Federal da Paraíba (UFPB) - João Pessoa, PB – Brasil.
http://lattes.cnpq.br/0221265788966967
E-mail: marckson.dci.ufpb@gmail.com

Luiz Gustavo de Sena Brandão Pessoa
Doutorando em Ciência da Informação pela Universidade Federal da Paraíba (UFPB) – PB -Brasil. Mestre em Ciências Contábeis pela Universidade de Brasília (UnB) – DF -Brasil. Professor da Universidade Federal da Paraíba (UFPB) - João Pessoa, PB - Brasil
http://lattes.cnpq.br/3978745023294083
E-mail: gustavobrandao@bol.com.br

Tereza Ludimila de Castro Cardoso
Mestranda em Ciência da Informação pela Universidade Federal da Paraíba (UFPB) – PB - Brasil. Especialização em Saúde Pública pela Universidade de São Paulo (USP) – SP - Brasil.
http://lattes.cnpq.br/8053283790009967
E-mail: luddyjampa@gmail.com

Submetido em: 02/12/2020. Aprovado em: 08/03/2021. Publicado em: 28/07/2021.

RESUMO

A tecnologia da informação trouxe um paradigma que se preocupa com a questão da disseminação dos dados que estão disponíveis nos diversos ambientes informacionais. É, nesse contexto, que a ciência de dados traz uma problemática para a Ciência da Informação: analisar os aspectos emergentes de tratamento, uso e reuso dos dados abertos a partir das necessidades informacionais dos usuários. Esse artigo propõe uma reflexão do paradigma dos dados e do tratamento que a Ciência da Informação pode fazer com a disponibilidade dos dados governamentais abertos. A proposta buscou verificar se os municípios que compõem a microrregião do litoral norte do Estado da Paraíba estão disponibilizando os dados para a sociedade, como preceitua a legislação de acesso à informação e transparência pública. A metodologia utilizada corresponde a uma pesquisa documental e descritiva, com tratamento de dados realizado por estatística simples através do aplicativo *LibreOffice Calc*. Foi utilizado o modelo de indicadores disponibilizados pelos relatórios da Controladoria Geral da União. Os resultados demonstram que os municípios estudados iniciaram o processo de implantação, mas que em alguns itens ainda precisam de atenção do gestor, principalmente no que se refere a disponibilizar o dado em tempo real.

Palavras-chave: Paradigma dos dados. Acesso à Informação. Transparência pública.

Application of Open Government Data to Information Science

ABSTRACT

The information technology has brought a paradigm that is concerned with the issue of the dissemination of data that are available in the various informational environments. It is in this context that data science brings a problem to the information science: to analyze the emerging aspects of treatment, use and reuse of data opened from the informational needs of users. This article proposes a reflection of the data and treatment paradigm that information science can do with the availability of open governmental data. The proposal sought to verify whether the municipalities that comprise the microregion of the northern coast of the state of Paraíba are making available the data to society, as preceded by the law of Access to information. The methodology used corresponds to a documental and descriptive research, with data processing performed by simple statistics through the LibreOffice CALC application. The model of indicators made available by the reports of the Comptroller General of the Union. The results show that the municipalities studied started the implantation process, but in some items still need the manager's attention, especially in terms of providing the data in real time.

Keywords*: Data paradigm. Access to information. Public transparency.*

Aplicación de Datos Gubernamentales Abiertos bajo la luz del a Ciencia de la Información

RESUMÉN

La tecnología de la información ha traído un paradigma relacionado con el tema de la difusión de datos disponibles en diferentes entornos de información. Es en este contexto que la ciencia de datos trae un problema a la Ciencia de la Información: analizar los aspectos emergentes del tratamiento, uso y reutilización de datos abiertos de las necesidades informativas de los usuarios. Este artículo propone un reflejo del paradigma de datos y el tratamiento que la ciencia de la información puede hacer con la disponibilidad de datos abiertos del gobierno. La propuesta buscaba verificar si los municipios que conforman la microrregión de la costa norte del Estado de Paraíba están poniendo los datos a disposición de la sociedad, según lo prescrito por la Ley de Acceso a la Información. La metodología utilizada corresponde a una investigación documental y descriptiva, con tratamiento de datos realizado mediante estadísticas simples a través de la aplicación LibreOffice Calc. Se utilizó el modelo de indicadores puesto a disposición por los informes del Contralor General de la Unión. Los resultados muestran que los municipios estudiados comenzaron el proceso de implementación, pero que en algunos ítems aún necesitan la atención del gerente, principalmente en lo que respecta a la disponibilidad de los datos en tiempo real.

Palabras clave: *Paradigma de datos. Acceso a la información. Transparencia pública.*

INTRODUÇÃO

No momento atual de complexidade econômica, política e social, a Ciência da Informação (CI), como uma área que estuda o fenômeno informação e seus aspectos funcionais e de tratamento de dados, faz-nos refletir sobre as necessidades de se pensar a transparência das contas públicas à luz das mudanças tecnológicas e informacionais.

A informação surge como uma fonte inesgotável de evolução voltada para uma sociedade cada vez mais tecnológica. Para Capurro e Hjorland (2007) este conceito, enquanto conhecimento surge no contexto de explosão tecnológica no período pós Segunda Guerra, no qual a informação desempenha um papel central para a sociedade. Para Capurro (2003), do ponto de vista epistemológico, a CI apresenta três tipos de paradigmas: o físico, o cognitivo e o social. O paradigma físico remete aos sistemas informatizados, sendo este fortemente influenciado pela questão tecnológica (ALMEIDA *et al,* 2007). Dessa forma, visa, prioritariamente, a uma "gestão de dados" mais eficiente, desenvolvendo e aperfeiçoando seus métodos.

Os paradigmas da CI se relacionam com os paradigmas científicos, trazendo para a atual conjuntura a significação necessária para a ciência, pois sem os sistemas informatizados, sem o usuário e sem a necessidade de informação de uma comunidade científica não haveria pesquisa de fato**.** Paralelamente a esse momento na CI, Jim Gray realiza, através de ferramentas computacionais, experimentos com o tratamento de grandes quantidades de dados disponibilizados a partir de outros cientistas das mais diferentes áreas, como observam Hey, Tansley e Tolle (2009), para eles, seria um momento novo para a história da ciência.

Considerando esse aspecto, o Brasil vem adotando algumas práticas de transparência, governança e *accountability* – termo que implica a responsabilidade do gestor em prestar contas à sociedade de suas decisões de aplicação de recursos públicos ou privados -, participação nas contas públicas nacionais, através de ferramentas de disponibilização de dados à sociedade, para que se possa acompanhar a gestão que é materializada pelas tomadas de decisões e condutas do gestor público. Assim, uma dessas ferramentas foi a implantação de uma política de abertura de dados públicos à sociedade, que pode ser considerado uma quebra de paradigma para os gestores.

Assim, desde a implantação da Lei de Acesso à Informação – LAI, o governo federal vem buscando uma série de medidas visando à disponibilização dos dados abertos governamentais, favorecendo políticas de transparência e *accoutability*. Neste sentido, os Tribunais de Contas têm um papel preponderante no acompanhamento dessa implantação, uma vez que esses órgãos fazem todo o acompanhamento dos dados disponibilizados das contas públicas da União, dos Estados e dos Municípios.

É, nesse contexto, que os dados abertos governamentais devem estar disponibilizados em seus portais em tempo real, assim como preceitua a LAI (BRASIL, 2011)**.** Dessa forma, o presente artigo busca realizar uma pesquisa com os municípios do estado da Paraíba com relação ao atendimento desse preceito legal a partir da seguinte questão de pesquisa: de que forma os municípios do estado da Paraíba estão disponibilizando para a sociedade os dados para o acompanhamento da gestão municipal?

Como objetivo geral pretende-se investigar se os municípios que compõem o Estado da Paraíba estão disponibilizando para a sociedade os dados para o acompanhamento da gestão pública municipal. Como objetivos específicos, temos:

- realizar uma busca nos portais institucionais dos municípios pesquisados;
- verificar, a partir de indicadores, se os dados pesquisados estão disponibilizados de acordo com a legislação de acesso à informação e transparência pública;
- tratar os dados pesquisados, analisando-os estatisticamente.

DADOS ABERTOS GOVERNAMENTAIS E SUA RELAÇÃO COM A CIÊNCIA DA INFORMAÇÃO

O termo e-Science foi introduzido por Jhon Taylor em 2001, de acordo com Sales, Souza e Sayão (2014). Este autor definiu e-Science como algo que mudaria a forma de fazer ciência, através de uma colaboração global em áreas chave da ciência e de uma subsequente geração de infraestrutura que possibilitaria esta colaboração. Dessa forma, o e-Science apresenta-se como o quarto paradigma científico.

A obra de Oliveira e Silva (2016) destaca também que o quarto paradigma científico trata de uma nova abordagem de comunicação científica, gerenciamento, curadoria, preservação com finalidade de colaboração mútua e acesso livre à publicação dos dados científicos. Esses aspectos demostram uma interação entre a Ciência da Informação e a Ciência de Dados, que como uma área interdisciplinar, buscando interagir técnicas para gerenciar, armazenar, recuperar e publicar dados.

Na visão de Harrison *et al.* (2012), são fundamentais e básicas, para o desenvolvimento de competências, as relações entre democracia, informação e transparência. Assim, a Ciência da Informação, bem como a Ciência de Dados é essencial para que essa busca pelo acesso democrático de dados abertos seja efetivamente realizada. O conceito de dados abertos, na visão de Sayão e Sales (2013), está associado à livre disponibilidade para o reuso em outras investigações científicas, possibilitando outros tratamentos, aplicações e resultados. Essa possibilidade requer uma reflexão com relação aos direitos de autoria, patentes e outros mecanismos de controle de autoria intelectual. Dados para serem abertos devem estar disponíveis para *downloads* gratuitos, com livre permissão para cópias, verificações, aplicações e demais tratamentos que gerem novas descobertas e possibilidades de uso.

No Brasil, a discussão com relação à abertura de dados governamentais não é recente. O texto constitucional prevê o direito da sociedade de receber informações dos órgãos públicos, dando à informação um caráter democrático e social. No entanto, não normatiza nem cria regras de quais informações são disponibilizadas nem dos meios. Ainda no que se refere à Administração Pública, o artigo 37º da CF/88 preceitua os princípios que a norteiam. Neste sentido, destacamos o princípio da Publicidade que trata da obrigatoriedade de tornar público os atos de todos os poderes (Executivo, Legislativo e Judiciário) e entes federados (União, Estados, Municípios e Distrito Federal).

A questão dos dados abertos também está relacionada ao princípio administrativo da Publicidade, requisito de eficácia dos atos administrativos de caráter informativo à sociedade, que também inspirou um dos pilares da Lei de Responsabilidade Fiscal – LRF/2000 (Lei Complementar 101/2000). A LRF traz em seu texto regras norteadoras da gestão pública, que a rigor a doutrina consagra como pilares da boa administração (BRASIL, 2000).

Nesse sentido, a ênfase da boa gestão pública é baseada no planejamento, responsabilização, controle e transparência. Este último coloca à disposição da sociedade, diversos meios de divulgação dos atos administrativos, através de relatórios orçamentários, financeiros e patrimoniais, que, de forma periódica, os entes são obrigados a disponibilizar. Neste sentido, as palavras de Carvalho (2009, p. 83) já retratavam o tipo de informação e os meios e ferramentas de comunicação disponíveis para a sociedade:

> A transparência dos atos públicos deve ser implementada por meio da divulgação nos meios de comunicação, inclusive eletrônico, dos resultados da gestão fiscal e das prestações de contas e pareceres prévios emitidos pelos tribunais de contas.

Percebe-se que os meios de comunicação utilizados tradicionalmente pela administração pública se resumiam ao rádio, televisão e jornal. Atualmente, pressupõem também a necessidade do uso da *internet,* uma vez que os portais institucionais são os canais mais rápidos e eficazes de comunicação entre a administração e a sociedade. Outro aspecto relevante é o tipo de informação que é requisitada, como os resultados da gestão fiscal, prestação de contas e pareceres de tribunais de contas.

Essas informações permitem à sociedade acompanhar o andamento da gestão e cobrar dos gestores ações para melhoria de desempenho orçamentário e financeiro, e, consequentemente, social.

Percebe-se que o advento da LAI tem por finalidade materializar o que já estava previsto na CF/88, quanto aos direitos de acesso à informação, uma vez que, com a LAI, foi possível descrever as regras e os meios que a informação deve ser disponibilizada pelos diversos órgãos que compõem a Administração Pública. Fica implícito que todos os atores envolvidos (sociedade e governo) devem se adequar ao regramento dessa lei.

Entretanto, o fato de a LAI ter sido criada com essa finalidade, não garante à sociedade sua plena eficácia já que "[...] as normas jurídicas são necessárias e fundamentais para a institucionalização de um campo de ação, porém são insuficientes para garantir a implementação efetiva de uma nova orientação da ação estatal e social" (GONZÁLEZ DE GÓMEZ, 1999 p. 69). Nesse aspecto, é necessário garantir, por parte da administração, a implantação de outros requisitos necessários à compreensão do que está sendo disponível a sociedade.

Neste cenário, cabe destacar a importância dos canais eletrônicos institucionais de serviços de informação e acesso ao cidadão que estão disponíveis desde a promulgação da LAI. De acordo com a Controladoria Geral da União – (CGU), o Serviço de Informação ao Cidadão – (SIC) "[...] permite que qualquer pessoa, física ou jurídica, encaminhe pedidos de acesso à informação, acompanhe o prazo e receba a resposta da solicitação realizada para órgãos e entidades do Executivo Federal. O cidadão ainda pode entrar com recursos e apresentar reclamações sem burocracia" (BRASIL, 2020, *online*).

Esse serviço está previsto no art. 9º da LAI e se consagra como um relevante elo entre a sociedade e a administração pública, uma vez que caracteriza formalmente a solicitação da informação em tempo real, e, ao mesmo tempo, obriga à administração pública responder com prazos o que foi solicitado.

Para que atinja seus objetivos, é necessário que a administração pública tenha não só a informação, mas também servidores capacitados para prestar a informação desejada.

Por parte da sociedade, é necessário que se busque conhecimentos sobre o tipo de dado que deseja obter, afim de que a formalização do pedido seja coerente com suas necessidades informacionais. No presente estudo, os conhecimentos necessários são relacionados às aplicações dos recursos que competem à gestão pública quanto às questões orçamentárias, financeiras e patrimoniais. Neste sentido, estão compreendidos os aspectos relativos à previsão e execução de receitas e fixação e execução de despesas, já preceituados pela LRF/2000.

Outro ponto que merece destaque é com relação às informações disponibilizadas nos portais institucionais sobre os procedimentos das licitações. Na visão de Mello (2004, p. 483), licitação é caracterizada por um certame em que entidades do governo promovem abertura de disputa entre interessados a manter relações patrimoniais com o governo, de forma que esta escolha a proposta mais vantajosa, de acordo com suas conveniências. Pressupõe entre aos que possuam as atribuições e aptidões necessárias à relação contratual, uma competição em igualdade de condições.

No Brasil, esse procedimento é previsto na Lei Federal nº 8.666/93 – Lei das Licitações e Contratos (BRASIL, 1993) e está prevista na CF/88 no art.37º, inciso XXI. Essa lei tem por finalidade estabelecer normas obrigatórias de contratação de serviços ou aquisições de bens, bem como procedimentos de alienações. Cabe ao ente que deseja contratar, estabelecer os critérios daquele objeto que pretende licitar.

Via de regra, salvo exceções previstas em lei, só devem existir relações contratuais de venda de produtos ou serviços de empresas com a administração pública através de processo licitatório.

Tais procedimentos são objeto do olhar da legislação de acesso e transparência pública, que obriga a disponibilização das informações nos portais institucionais, e devem ser acompanhados e refletidos continuamente pelo cidadão, uma vez que estão diretamente relacionados com a prestação de serviço e aquisição de bens que serão utilizados pela sociedade.

No que se refere à temporalidade que a legislação de acesso e transparência pública obriga os portais a disponibilizarem informações à sociedade, cabe destaque ao aspecto destas serem disponibilizadas em "tempo real". Nesse aspecto, a tecnologia da informação tem um papel preponderante uma vez que possibilita a utilização de ferramentas tecnológicas para que o governo possa utilizar em favor do acesso à informação pela sociedade. O Decreto nº 7.185/2010 dispõe sobre o padrão mínimo de qualidade do sistema integrado de administração financeira e controle. Nesse decreto é previsto que a disponibilização por meio eletrônico da informação em tempo real:

> [...] possibilite amplo acesso público, até o primeiro dia útil subseqüente à data do registro contábil no respectivo SISTEMA, sem prejuízo do desempenho e da preservação das rotinas de segurança operacional necessários ao seu pleno funcionamento (BRASIL, 2010, *online*).

Esse procedimento vem sendo objeto de discussão quando se trata de acompanhamento da gestão pública, isso porque muitas vezes os gestores não observam essa determinação legal de forma satisfatória. Nesse aspecto, o papel da sociedade e dos órgãos fiscalizadores é essencial para acompanhar a aplicação desse dispositivo, já que a forma de disponibilizar em tempo real tem relação com a tempestividade do ato administrativo.

METODOLOGIA

A presente pesquisa é resultado de um projeto de Extensão Universitária em andamento vinculado ao Campus IV da Universidade Federal da Paraíba. A proposta do projeto busca investigar se os 223 municípios do Estado da Paraíba estão disponibilizando, de acordo com a legislação de acesso à informação e transparência pública, as informações em seus portais governamentais. O Estado da Paraíba possui 4 mesorregiões subdivididas em 23 microrregiões. Para este trabalho, serão apresentados resultados obtidos a partir dos dados de 3 microrregiões, correspondendo a um total de 21 municípios, assim subdivididos:

- 11 municípios que compõem a microrregião do litoral paraibano, composta pelos municípios de Rio Tinto, Mamanguape, Mataraca, Marcação, Curral de Cima, Pedro Régis, Itapororoca, Jacaraú, Curral de Cima, Baia da Traição e Capim;
- 04 municípios que compõem a microrregião do litoral sul paraibano, composta pelos municípios de Alhandra, Caaporã, Pedras de Fogo e Pitimbu;
- 06 municípios que compõem a microrregião de João Pessoa, composta pelos municípios de João Pessoa, Cabedelo, Lucena, Bayeux, Santa Rita e Conde.

Esta pesquisa é caracterizada como documental e quanti-qualitativa. Com a finalidade de alcançar o objetivo proposto, foi considerado o acesso digital aos portais institucionais dos municípios que compõem as microrregiões do Litoral Norte, do Litoral Sul e de João Pessoa, no Estado da Paraíba. O critério de escolha amostral desses municípios foi por conveniência, em função de estes estarem localizados em microrregiões geográficas próximas da localização do Campus IV da Universidade Federal da Paraíba.

Quanto aos dados que foram verificados, utilizou-se como referência, o Modelo de Indicadores de Verificação para Avaliação dos Portais da Transparência, que são disponibilizados na apresentação dos relatórios da Controladoria Geral da União.

Neste, estão contidos os indicadores com as categorias de pesquisa mais relevantes no processo de adequação de transparência e *accoutability*. O modelo categoriza o assunto de item e questiona se o dado está disponível ou não, no portal institucional do município.

Os dez indicadores utilizados na pesquisa foram elencados a partir da base legal da LAI, que buscava informação de dados quanto a:

1. Regulamentação da LAI;
2. Implantação de serviço de informação ao cidadão (SIC);
3. Alternativa de enviar pedido de forma eletrônica ao SIC;
4. Previsão e arrecadação de receitas;
5. Empenho, liquidação e pagamento de despesas;
6. Classificação orçamentária da unidade que financiou o gasto;
7. Pessoa física ou jurídica beneficiada com o pagamento;
8. Indicação de procedimento licitatório;
9. Informação sobre prestação do serviço e entrega de produto;
10. Atendimento ao requisito "tempo real" (resultados específicos são apresentados na tabela 2, em virtude de sua especificidade com relação aos indicadores 1 a 9).

O tratamento dos dados foi feito através de planilha eletrônica do *LibreOffice* e os campos foram preenchidos de forma codificada ao atendimento ou não da LAI, onde "S" corresponde à resposta afirmativa e "N" corresponde à resposta negativa. Foi feita uma estatística simples e descritiva no tratamento dos dados.

DISCUSSÃO DOS RESULTADOS

A partir dos dados obtidos na pesquisa realizada nos portais governamentais dos municípios listados anteriormente, foi verificado que todos os municípios pesquisados possuem aspectos satisfatórios de transparência em seus portais, o que viabilizou a pesquisa. Posteriormente, em todos os portais, foram realizadas buscas nos *links* que tratavam dos itens dos dados abertos, verificando-se que a maioria possuía o *link* "Transparência Fiscal" e "Sistema de Informação ao Cidadão - SIC" como metadados. Observou-se ainda que não há uma padronização de procedimentos, preservando as especificidades de cada portal, sem comprometer a essência dos dados requisitados na LAI.

Dessa forma, o acesso ao *link* possibilitou verificar, a partir da pesquisa de cada indicador, se o portal disponibiliza a informação desejada conforme previamente estabelecido na LAI. O tratamento dos dados está demonstrado na tabela 1 e na tabela 2. A tabela 1 se refere aos indicadores 1-9; e a tabela 2 refere-se aos indicadores 1-9 em relação ao indicador 10 que considera as ações realizadas em tempo real.

- Indicador 01: quanto à regulamentação da LAI – a regulamentação da LAI é o ato do gestor que regulamenta, no âmbito do seu município, essa lei. Inicialmente, foi verificado se nos portais dos municípios havia esse ato normativo.

A pesquisa demonstrou que 76,1% dos municípios pesquisados haviam feito a regulamentação, que normalmente é realizada através de um ato de decreto municipal. Neste sentido, o acompanhamento dos Tribunais de Contas vem sendo realizado de forma constante, e, uma vez regulamentada a lei, o município sistematiza todo o processo de sua implantação. Observou-se que em 23,9% dos municípios não constavam essa regulamentação por parte do gestor municipal, entretanto havia a citação da Lei de Acesso à Informação como referência para alimentar seus portais. Ao relacionar esse indicador com outro que prevê a situação "tempo real" (tabela 2), observou-se que 19% não disponibilizaram esse decreto no momento da pesquisa, apenas indicavam qual foi o decreto que o regulamentou.

- Indicado 2: quanto à implantação do SIC – esse item questionou se o município implantou o SIC. Observou-se que 100% dos municípios haviam implantado esse serviço, que possibilita ao cidadão buscar de forma presencial o dado que deseja.

Observou-se também que grande parte dos portais disponibilizam também formulários para que o cidadão preencha, imprima e protocole o pedido no órgão correspondente. Ao relacionar esse indicador com o outro que trata da situação em tempo real (tabela 2), observou-se a mesma situação de adequação da totalidade dos municípios a esse quesito.

- Indicador 3: quanto à disponibilização de pedido eletrônico do SIC - esse indicador verifica se existe a alternativa de se buscar a informação de forma eletrônica. Dessa forma, 95,2% dos municípios pesquisados já observam essa determinação legal.

Como não há uma padronização, uns solicitam um cadastro prévio do requerente, solicitando inclusive dados pessoais, já outros pedem apenas o correio eletrônico para envio da informação. No único município que não disponibiliza essa informação, percebeu-se que existe o link, mas ao buscar não se consegue acesso. Ao relacioná-lo com o indicador que trata do tempo real (tabela 2), 14,3% dos municípios apresentavam algum tipo de inconsistência na plataforma eletrônica, o que comprometeu a disponibilidade da informação, e, consequentemente, baixou o índice da pesquisa nesse indicador.

- Indicador 4: quanto à apresentação de previsão e arrecadação de receitas – esse item busca verificar se o município disponibiliza relatórios que demonstrem a situação legal quanto à execução das receitas. Cabe destacar que a previsão das receitas ocorre anteriormente a sua arrecadação, e a arrecadação é realizada à medida que os agentes arrecadadores recebem os recursos.

Neste aspecto, observou-se que 100% dos municípios se adequaram a esse ponto. Na realidade, mesmo antes da implantação da LAI, esse item já era de disponibilização obrigatória, uma vez que a LRF/00 já regulamentava e o próprio Tribunal de Contas, como entidade de controle externo, já vem fiscalizando se o município vem demonstrando periodicamente a situação das Receitas. Entretanto, quando se verifica esse indicador relacionado com o quesito "tempo real" observou-se que os municípios que disponibilizam esses dados correspondem a 85,7%, restando 14,3% que não disponibilizam dados em tempo real - conforme a tabela 2, comprometendo a atualização dos dados.

- Indicador 5: quanto ao empenho, liquidação e pagamento da despesa - em ato contínuo à pesquisa do indicador das receitas, o indicador que trata da realização dos empenhos, liquidação e pagamentos das despesas também foi avaliado.

Observou-se que os municípios demonstram que estão realizando de forma sistemática e rotineira os atos de empenhos, liquidação e pagamentos das despesas. Assim como as receitas, as despesas também são objetos de transparência já previstos na LRF/00, o que explica a adequação a este item em 100% dos municípios pesquisados. Cabe destacar a ocorrência de que 85,7% dos municípios estão adequados quando se busca a informação em tempo real (tabela 2) e 14,3% estão em situação de dados desatualizados.

- Indicador 6: quanto à unidade que realizou o gasto – esse indicador busca verificar se há a informação das unidades orçamentárias que realizaram o gasto. Os resultados demonstram que em 90,4% há a informação da unidade que realizou o gasto e, em 9,6% dos municípios pesquisados, não estão disponibilizando as unidades que estão executando gastos.

Essas unidades também estão relacionadas com a gestão dos convênios celebrados com o Governo Federal, nas áreas da Saúde e Educação e/ou outros entes financiadores do gasto público.

Assim, ao relacionar com o indicador "tempo real", verificou-se que o índice dos municípios que demonstram atender esse quesito, decresceu para 80,1%, aumentando para 19,9% os municípios que não atendem à informação da unidade que realizou o gasto em tempo real (tabela 2).

- Indicador 7: quanto à pessoa física ou jurídica beneficiada com pagamento - ainda em função dos pagamentos realizados pelos municípios, esse indicador aponta se os recursos pagos são recebidos por pessoa física ou pessoa jurídica.

Neste quesito, observou-se que 38% dos municípios pesquisados não apresentam os dados, correspondendo aos municípios de Mamanguape, Itapororoca, Pedro Régis, Marcação, Cuité de Mamanguape, Santa Rita e Cabedelo. Esse dado é imprescindível para que a sociedade acompanhe quem está recebendo os recursos, interferindo na análise da *accountability* na gestão municipal e fica ainda mais comprometido quando analisado em relação ao indicador "tempo real" (tabela 2), demonstrando que 61,9% dos municípios não apresentam a informação de forma atualizada.

- Indicador 8: indicação de existência de procedimento licitatório – o procedimento licitatório é alvo de acompanhamento e fiscalização dos órgãos de controle e também merece atenção da sociedade.

É, nesse item, que estão demonstradas as intenções de compra e prestação de serviços que o Estado pretende realizar com o particular, conforme os ditames da Lei nº 8.666/92. O portal precisa apresentar informações quanto à modalidade, tipo, editais e demais regras do processo de escolha de quem vai prestar o serviço ou vender o produto ao Estado. A LRF também contempla regras de transparência nesse processo. Na pesquisa aos portais, observou-se que em 14,3% dos casos, os municípios não apresentam em seus portais os dados de procedimentos licitatórios realizados no âmbito da gestão. Quando verificados, estes, em relação ao quesito "tempo real", observou-se que dos que não apresentam dados de licitação aumentam para 47,7%.

- Indicador 9: informação sobre prestação de serviço e/ou entrega do produto – esse indicador relaciona-se com a despesa na etapa da liquidação, busca também acompanhar as relações contratuais com relação ao comprometimento do contratante e à qualidade do serviço e do produto.

Durante a pesquisa, foi possível observar que 71,4% dos municípios pesquisados apresentam informações sobre prestação de serviço e/ou entrega de produto, entretanto, 28,6% dos municípios ainda estão deficientes nesse quesito. Ao relacionar este item ao indicador referente ao tempo real, observou-se que, em 47,7% dos casos, os municípios não apresentam informações sobre esse indicador.

- Indicador 10: atendimento ao requisito "tempo real" - esse indicador está apresentado na tabela 2 e foi relacionado em todas as análises anteriores, buscando verificar se os itens relacionados aos indicadores 1 a 9 estavam disponibilizados em tempo real.

Para que esse indicador seja validado, é necessário que as informações estejam atualizadas de acordo com o momento em que o fato gerador ocorra. Para tanto, faz-se necessário que o município tenha uma equipe que trabalhe a tecnologia da informação de forma síncrona e otimizada com o controle interno. Nesse contexto, esse indicador pode demonstrar limitações nos portais pesquisados, principalmente em função de eventuais problemas técnicos dependentes da tecnologia da informação.

A pesquisa mostra uma situação específica, na qual os atores podem pesquisar dados governamentais para acompanhamento da gestão pública, em que a Ciência da Informação, assim como a Ciência de Dados em função da interdisciplinaridade própria da sua atuação, pode auxiliá-los na busca de soluções para as diversas questões que permeiam a sociedade, quanto às respostas da atuação do gestor público. Percebe-se, então, uma aplicação dos conceitos de dados abertos, propostos por Sayão e Sales (2013), associando-os à livre disponibilidade para o reuso e possibilidades para novos tratamentos, aplicações e resultados.

Tabela 1 – Dados abertos nos portais dos municípios das microrregiões do Litoral Norte, João Pessoa e Litoral Sul do Estado da Paraíba

Relação dos municípios da microrregião do litoral sul e litoral norte do Estado da Paraíba	1.Regulamentação LAI	2.Implementação do SIC	3.Pedido Eletrônico do SIC	4.Previsão e arrecadação de receitas	5.Empenho e pagamento da despesa	6.Unidade que financiou o gasto	7.PF ou PF beneficiária do pagamento	8.Indicação de Procedimento Licitatório	9.Prest. serviço ou inf.de entrega do bem
Baía da Traição	S	S	S	S	S	S	S	S	S
Capim	S	S	S	S	S	S	S	S	S
Cuité de Mamanguape	S	S	S	S	S	S	N	N	S
Curral de Cima	S	S	S	S	S	S	S	S	S
Itapororoca	S	S	S	S	S	S	N	S	S
Jacaraú	S	S	S	S	S	S	S	S	S
Mamanguape	S	S	S	S	S	S	N	S	N
Marcação	S	S	S	S	S	S	N	S	S
Mataraca	S	S	S	S	S	S	S	S	N
Pedro Régis	S	S	S	S	S	S	N	S	S
Rio Tinto	S	S	S	S	S	S	S	S	N
Alhandra	S	S	S	S	S	N	N	N	N
Caaporã	N	S	S	S	S	S	S	S	S
Pedras de Fogo	S	S	S	S	S	S	S	S	S
Pitimbu	S	S	S	S	S	S	S	S	S
Bayeux	N	S	N	S	S	S	S	S	S
Cabedelo	N	S	S	S	S	N	N	N	N
Conde	S	S	S	S	S	S	S	S	S
Joao Pessoa	S	S	S	S	S	S	S	S	S
Lucena	N	S	S	S	S	S	S	S	S
Santa Rita	N	S	S	S	S	S	N	S	N
SIM	76,1%	100%	95,2%	100%	100%	90,4%	62,0%	85,7%	71,4%
NÃO	23,9%	0%	4,8%	0%	0%	9,6%	38,0%	14,3%	28,6%

Fonte: Dados da pesquisa (2020).

Tabela 2 − Demonstração dos dados abertos nos portais dos municípios das microrregiões do Litoral Norte, João Pessoa e Litoral Sul do Estado da Paraíba que atendem ao requisito "tempo real", conforme indicador 10

Relação dos municípios da microrregião do litoral sul e litoral norte do Estado da Paraíba	1.Regulamentação LAI	2.Implementação do SIC	3.Pedido Eletrônico do SIC	4.Previsão e arrecadação de receitas	5.Empenho e pagamento da despesa	6.Unidade que financiou o gasto	7.PF ou PF beneficiária do pagamento	8.Indicação de Procedimento Licitatório	9.Prest. serviço ou inf.de entrega do bem
Cuité de Mamanguape	S	S	S	S	S	S	N	N	S
Marcação	S	S	S	S	S	S	N	N	N
Pedro Régis	S	S	S	S	S	S	N	S	S
Jacaraú	S	S	S	S	S	S	N	S	N
Curral de Cima	S	S	S	S	S	S	S	S	S
Baía da Traição	S	S	S	S	S	S	S	S	S
Itapororoca	S	S	S	S	S	S	N	S	N
Mataraca	S	S	S	S	S	S	N	S	N
Capim	S	S	S	S	S	S	S	S	S
Rio Tinto	S	S	S	S	S	S	S	N	N
Mamanguape	S	S	S	S	S	S	N	N	N
Alhandra	S	S	N	S	S	N	N	N	N
Caaporã	S	S	S	S	S	S	S	S	S
Pedras de Fogo	S	S	S	N	N	N	N	N	N
Pitimbu	S	S	S	S	S	S	S	S	S
Bayeux	N	N	N	S	S	S	S	S	S
Cabedelo	N	S	S	N	N	N	N	N	N
Conde	S	S	S	N	N	N	N	N	N
Joao Pessoa	S	S	S	S	S	S	S	S	S
Lucena	N	S	S	S	S	S	N	N	S
Santa Rita	N	N	N	S	S	S	N	N	N
SIM	81,0%	100%	85,7%	85,7%	85,7%	80,1%	38,1%	52,3%	47,7%
NÃO	19,0%	0%	14,3%	14,3%	14,3%	19.9%	61,9%	47,7%	52,3%

Fonte: Dados da pesquisa (2020).

CONSIDERAÇÕES FINAIS

A situação paradigmática do que se entende hoje por Ciência de Dados demonstra que a Ciência da Informação está diante de um novo cenário que gera novos desafios. A pesquisa em tela demonstra, através dos dados obtidos, que não basta apenas uma legislação para que o gestor disponibilize os dados, são necessárias outras variáveis para que a informação alcance sua finalidade social. Além do acompanhamento dos órgãos de controle e da sociedade, são necessárias também competências específicas para a geração da informação, a partir dos dados disponibilizados.

Durante o processo da pesquisa, foi possível constatar que a grande maioria dos municípios pesquisados regulamentou a LAI e implantou o SAC, tanto de forma presencial quanto de forma eletrônica. Entretanto, cabe destacar que, em 14,3% dos municípios, esta última forma de SAC ainda está sendo disponível de forma incipiente.

Outro ponto que merece destaque na pesquisa aos portais é a consolidação da disponibilização dos dados que já era prevista desde a LRF/00; que a LAI também traz em seu texto, a exemplo dos itens relacionados à previsão e arrecadação de receitas, informações sobre empenho, liquidação e pagamento de despesas, além de informações sobre procedimentos licitatórios. Esses itens, na grande maioria dos municípios pesquisados, estavam sendo observados. Entretanto, não se pode deixar de afirmar que, ao relacionar esses indicadores ao quesito "tempo real", constatou-se que, em muitos casos, o item pesquisado estava disponível no portal, mas era insuficiente, já que havia necessidade de ser atualizado.

Por fim, para o que esta pesquisa se propôs a realizar, podemos afirmar que seus objetivos foram atingidos, sendo possível constatar que o principal ponto de preocupação está relacionado com a questão do atendimento da disponibilização dos dados em tempo real. Essa indisponibilidade traz limitações ao trabalho do pesquisador, bem como daquele que pretende acompanhar a gestão.

Outra questão de limitação relevante é com relação ao momento (lapso temporal) em que a pesquisa foi realizada e os dados foram alimentados por parte da equipe de TI do município; questões técnicas dos portais e limitações específicas de quadro técnico qualificado de servidores de cada município. Tais considerações levam o pesquisador a refletir também sobre o alcance dos propósitos da abertura dos dados no Brasil.

REFERÊNCIAS

ALMEIDA, D. P. R. *et al.* Paradigmas Contemporâneos da Ciência da Informação: a recuperação da informação como ponto focal. *Revista Eletrônica Informação e Cognição,* Marília, v.6, p.16-27, 2007. Disponível em: http://www.brapci.inf.br/_repositorio/2010/03/pdf_fc4f01292e_0008415.pdf. Acesso em: 28 ago. 2019

BRASIL. Constituição (1988). *Constituição da República Federativa do Brasil.* Brasília, DF: Senado Federal: Centro Gráfico, 1988, 292 p. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 25 abr. 2020.

BRASIL. *Decreto nº 7.185*, de 27 de maio de 2010. Dispõe sobre o padrão mínimo de qualidade do sistema integrado de administração financeira e controle, no âmbito de cada ente da Federação, nos termos do art. 48, parágrafo único, inciso III, da Lei Complementar no 101, de 4 de maio de 2000, e dá outras providências. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2010/decreto/d7185.htm. Acesso em: 25 abr. 2020.

BRASIL. *e-SIC*. Sistema Eletrônico do Serviço de Informação ao Cidadão. 2020. Disponível em: https://esic.cgu.gov.br/sistema/site/index.aspx. Acesso em: 25 abr. 2020.

BRASIL. *Lei Complementar nº 101*, de 4 de maio de 2000. Estabelece normas de finanças públicas voltadas para a responsabilidade na gestão fiscal e dá outras providências. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/lcp/lcp101.htm. Acesso em: 25 abr. 2020.

BRASIL. *Lei nº 8.666*, de 21de junho de 1993. Regulamenta o art. 37, inciso XXI, da Constituição Federal, institui normas para licitações e contratos da Administração Pública e dá outras providências. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/L8666compilado.htm. Acesso em: 25 abr. 2020.

BRASIL. *Lei nº 12.527*, de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 25 abr. 2020.

CARVALHO, D. *Orçamento e Contabilidade Pública.* 6. ed. Rio de Janeiro: Elsevier, 2014.

CAPURRO, R. Epistemologia e ciência da informação. In: ENANCIB, 5., 2003. Belo Horizonte. *Anais*... Belo Horizonte: UFMG, 2003. Disponível em: http://www.capurro.de/enancib_p.htm. Acesso em: 28 ago. 2019.

CAPURRO, R.; HJORLAND, B. O. O conceito de informação. *Perspectivas em Ciência da Informação*, Belo Horizonte, v.1, n.1, p.148-207, 2007. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/54. Acesso em: 28 ago. 2019.

GONZÁLEZ DE GÓMEZ, M. N. Da política de informação ao papel da informação na política contemporânea. *Revista Internacional de Estudos Políticos,* Rio de Janeiro: UERJ/NUSEG, v. 1, n. 1, p. 67-93, abr. 1999.

HARRISON, T. M. *et al.* Open government and e-government: democratic challenges from a public value perspective. *Information Polity*, v. 17, n. 2, p. 83-97, 2012.

HEY, T.; TANSLEY, S.; TOLLE, K. (Ed.). Jim Gray on eScience: a transformed scientific method. In: HEY, T.; TANSLEY, S.; TOLLE, K. (Ed.). *The fourth paradigm: dataintensive scientific discovery*. Redmond: Microsoft Research, 2009. p. xvii-xxxi. Disponível em: http://digital.library.unt.edu/ark:/67531/metadc31516/. Acesso em: 31 ago. 2019.

MELLO, C. A. B. *Curso de Direito Administrativo*. 17. ed. rev. e atual. São Paulo: Malheiros, 2004.

OLIVEIRA, A. C. S; SILVA, E. M. Ciência aberta: dimensões para um novo fazer científico. *Informação & Informação,* Londrina, v. 21, n. 2, p. 5-39, maio/ago., 2016.

SAYÃO, L. F.; SALES, L. F. Dados de pesquisa: contribuição para o estabelecimento de um modelo de curadoria digital para o país. *Tendências da Pesquisa Brasileira em Ciência da Informação*, Belo Horizonte, v. 6, n. 1, 2013. Disponível em: http://inseer.ibict.br/ancib/index.php/tpbci/article/viewArticle/102. Acesso em: 04 set. 2019.

SALES, L. F.; SOUZA, R.F.; SAYÃO, L.F. Publicação Ampliada: um novo modelo de publicação científica voltada para os desafios de uma ciência orientada por dados. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15., 2014, Belo Horizonte. *Anais*... Belo Horizonte: ECI/UFMG, 2014. p.3471-3492. Disponível em: http://enancib2014.eci.ufmg.br/documentos/anais/anais-gt7/view. Acesso em: 28 ago. 2019.

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

Explorando a reconciliação de dados culturais na Wikidata: experimento aplicado com o acervo museológico do Museu Histórico Nacional

Luis Felipe Rosa de Oliveira
Doutorando em Ciência da Informação pela Universidade de Brasília (UnB) - Brasília, DF - Brasil. Mestre em Comunicação pela Universidade Federal de Goiás (UFG) – GO - Brasil.
http://lattes.cnpq.br/6498992926514286
E-mail:luisfelipeprf@gmail.com

Dalton Lopes Martins
Pós-Doutorado pela Universidade de São Paulo (USP) – SP - Brasil. Doutor em Ciência da Informação pela Universidade de São Paulo (USP) - São Paulo, SP - Brasil. Professor da Universidade de Brasília (UnB) - Brasília, DF - Brasil.
http://lattes.cnpq.br/3774617443225038
E-mail:dmartins@gmail.com

Submetido em: 22/09/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021.

RESUMO

Este estudo foi desenvolvido sob a perspectiva da web semântica e dos dados abertos ligados, com enfoque na técnica de reconciliação de dados e busca entender como ocorre o processo de reconciliação de dados culturais com a Wikidata, através da programação de scripts na linguagem Python, com o objetivo de contribuir para o entendimento de como se dá a aplicação de uma técnica de enriquecimento semântico em bases de dados culturais. Como metodologia, são descritas as etapas de desenvolvimento dos scripts de reconciliação de dados. E, como resultados, são apresentados os produtos da aplicação dos scripts na reconciliação de parte dos dados do acervo museológico do Museu Histórico Nacional com os objetos digitais da Wikidata. Chega-se à conclusão de que o processo de descrição do desenvolvimento dos scripts permitiu compreender melhor como ocorre a reconciliação de dados em acervos culturais, de que se deve dar mais atenção à normalização dos dados do acervo, e de que esse tipo de aplicação amplia o potencial de socialização do conhecimento em rede.

Palavras-chave: Web semântica. Dados abertos ligados. Acervos culturais. Enriquecimento semântico.

Exploring the reconciliation of cultural data on wikidata: experiment applied with the museum collection of the national historical museum

ABSTRACT

This study was developed from the perspective of the semantic web and the linked open data, focusing on the data reconciliation technique. It seeks to understand how the process of reconciling cultural data with Wikidata *occurs through the programming of scripts in the Python language, with the aim of contributing to the understanding of how to apply a technique of semantic enrichment in cultural databases. As a methodology, the stages of the data reconciliation scripts development are described. And as results, the products of the scripts application are presented in the reconciliation of part of the data of National Historical Museum museological collection with the digital objects of* Wikidata. *It is concluded that the process of describing the scripts development allowed to better understand how data reconciliation occurs in cultural collections, and that more attention should be paid to the normalization of the collection data, and that this type of application expands the potential for networked socialization of knowledge.*

Keywords*: Semantic web. Linked open data. Cultural collections. Semantic enrichment.*

Explorando la reconciliación de datos culturales en wikidata: experimento aplicado con la colección del museo del museo histórico nacional

RESUMEN

Este estudio se desarrolló desde la perspectiva de la web semántica y los datos abiertos vinculados, centrándose en la técnica de conciliación de datos. Busca comprender cómo ocurre el proceso de conciliación de datos culturales con Wikidata *a través de la programación de scripts en el lenguaje Python, para contribuir a la comprensión de la aplicación de una técnica de enriquecimiento semántico en las bases de datos culturales. Como metodología, se describen las etapas de desarrollo de los scripts de reconciliación de datos. Y como resultado, los productos de la aplicación de los scripts se presentan en la conciliación de parte de los datos de la colección museológica del Museo Histórico Nacional con los objetos digitales de* Wikidata. *Se concluye que el proceso de describir el desarrollo de los scripts nos permitió comprender mejor cómo se produce la conciliación de datos en las colecciones culturales, y que se debe prestar más atención a la normalización de los datos recopilados, y que este tipo de aplicación expande un potencial de socialización del conocimiento en red.*

Palabras clave*: Web Semántica. Datos abiertos vinculados. Colecciones culturales. Enriquecimiento semántico.*

INTRODUÇÃO

A apropriação crescente dos conceitos da web semântica pelos estudos em Ciência da Informação é notável, uma temática que consolida estudos importantes da área, como construção e aplicação de ontologias, estruturas de dados semânticos e interoperabilidade de bases de dados, contextualizando-os no ambiente digital da web, agregando as condições do fenômeno de interação em rede e a alta demanda pela capacidade de lidar com o volume e a variabilidade de dados gerados a cada segundo.

Se aventurar pelo universo de possibilidades da web semântica é tentador a qualquer profissional da informação, porém esse processo pode ser um pouco frustrante em um primeiro momento. Entender quais os aspectos da web semântica podem beneficiar suas atividades é um desafio, dado que é necessário convergir diferentes técnicas e estruturas agregadas em um ambiente dinâmico de conexão e reúso de informações para a criação de produtos e serviços informacionais. É, a partir desse sentimento, que este estudo se origina da necessidade de entender como processos inerentes da web semântica se dão, e como eles podem beneficiar o acesso e reúso da informação.

Vale elencar aqui o estudo desenvolvido por Santarém Segundo (2014), que apresentou uma exploração do uso do protocolo SPARQL (protocolo para consultas e manipulação de dados estruturados semanticamente) para recuperação da informação em bases semânticas, resultando na indicação da viabilidade do protocolo para acesso a esse tipo de conteúdo semântico na web, fundamentando a aplicação efetivada no estudo aqui proposto sob o contexto das informações de cunho cultural na web.

Mais especificamente, como acervos digitais de instituições de patrimônio cultural podem incorporar os benefícios e serem enriquecidos com os produtos de aplicações semânticas. O foco da pesquisa é demonstrar meios práticos e operacionais de se valer de bases de conhecimento já existentes para efetivar a ligação de dados e ampliar a possibilidade de adoção desse recurso para ampliar e enriquecer as fontes de informação culturais.

O movimento dessas instituições de patrimônio cultural de digitalização dos objetos culturais e a disponibilização online de seus acervos entre outras consequências condicionam uma dinâmica de difusão cultural em rede, com um grande potencial de consumo e agregação com outros provedores de dados. Enxergando que tópicos da web semântica podem se valer desse potencial, e que os profissionais atuais da Ciência da Informação, com conhecimento técnico prévio, podem desenvolver recursos. Nessa perspectiva, este estudo busca entender como se dá o processo de reconciliação de dados culturais com a Wikidata através da programação de scripts na linguagem Python.

Como objeto prático do estudo, foi utilizada a base de dados do acervo museológico disponível no repositório digital do Museu Histórico Nacional (MHN) disponível para acesso no link http://mhn.acervos.museus.gov.br/reserva-tecnica/. Este acervo faz parte de um conjunto de outros acervos publicados atualmente na web, ressaltando, aqui, a iniciativa do Instituto Brasileiro de Museus em promover a disponibilização on-line dos acervos digitais dos museus sob sua gestão através do software Tainacan[1]. Esse tipo de iniciativa reforça a estruturação de um conjunto de objetos digitais culturais disponíveis na web, abrindo espaço para o alto potencial do uso de aplicações semânticas para enriquecimento de acervos na área cultural.

Desse modo, o estudo aqui proposto se estrutura sob o desenvolvimento de scripts de reconciliação de dados com o objetivo de reconciliar parte do conjunto de dados do acervo museológico do MHN com objetos digitais da Wikidata, colocando esse acervo no universo dos dados ligados, e abrindo portas para o enriquecimento de seus dados.

Este experimento foi realizado com objetivo principal de contribuir para o entendimento de como se dá a aplicação de uma técnica de enriquecimento semântico em bases de dados culturais.

[1] Tainacan - https://tainacan.org/ Acesso em 19/09/2020.

Vale ressaltar ainda que este resultado é parte de um esforço maior de pesquisa e faz parte de um projeto de doutorado em andamento, que tem por objetivo implementar um serviço de reconciliação de dados para todos os repositórios digitais disponibilizados na plataforma Tainacan no âmbito do Instituto Brasileiro de Museus.

O artigo apresenta a seguir uma contextualização dos conceitos que fundamentam a aplicação do estudo, a metodologia com a descrição do desenvolvimento dos scripts de reconciliação de dados, e, em seguida, **a análise dos resultados aponta sínteses dos produtos dos scripts e alguma**s reflexões sobre os resultados. Por fim, nas considerações finais, **são discutidos os principais pontos alcançados e** as propostas de pesquisas futuras para continuar a contribuição do entendimento da aplicação deste tipo de técnica de enriquecimento semântico em acervos digitais culturais.

WEB SEMÂNTICA E RECONCILIAÇÃO DE DADOS

A reconciliação de dados é o conceito que fundamenta a pesquisa aplicada neste estudo, porém, antes de apresentar sua definição, é importante contextualizar os conceitos de web semântica e dados abertos ligados que fundamentam a temática sob a qual este artigo foi desenvolvido.

Idealizada por Tim Berners-Lee, a web semântica é uma "extensão da web", que possibilita um fluxo de acesso à informação mais estruturado para a leitura por softwares, potencializando funcionalidades de busca e reúso da informação, o que, consequentemente, afeta a forma como os usuários acessam a informação na web, amplificando as formas de como o conteúdo pode ser gerado e consumido em rede (BERNERS-LEE; HENDLER; LASSILA, 2001).

Essa visão da web semântica na prática é mais perceptível do ponto de vista do desenvolvimento de tecnologias de acesso e reúso da informação, pois a estruturação dos dados de forma semântica na web permitirá que repositórios digitais relacionem informações de diversos provedores digitais, como por exemplo, um acervo museológico publicado através do Tainacan na web poder ter seus dados de autoria relacionados com os dados de um provedor de controle autoridades como o VIAF[2], ou ainda dados de localização com um provedor de dados geográficos como o GeoNames[3].

Os resultados desse tipo de relacionamento dos dados são muitos. Da perspectiva da qualidade da informação, esse tipo de aplicação reduz a incerteza sobre os dados, pois ao conectá-los com um provedor com controle semântico, os dados são contextualizados e recebem identificadores únicos na web. Outra perspectiva é dada a partir da geração automática de conteúdo, uma vez que os dados estão estruturados de forma semântica, é possível gerar conteúdos de maneira automática, como são feitos os painéis de informação do Google que aparecem ao se pesquisar por uma entidade importante, por exemplo, quando se pesquisa por Tim Berners-Lee no Google[4], aparece em destaque um conjunto de fotos dele, um breve resumo biográfico obtido da Wikipedia , os livros publicados por ele e algumas pesquisas relacionadas.

Para que essas possibilidades semânticas existam de maneira mais frequente e acessível, é necessário que alguns padrões, que são indicados principalmente pelo W3C (*World Wide Web Consortium*), sejam adotados e aplicados, como o RDF (*Resource Description Framework*), que permite descrever objetos digitais a partir de uma estrutura semântica de triplas, envolvendo um sujeito, um predicado e um objeto. Vale ressaltar ainda o papel importante das ontologias e dos padrões de metadados, que, quando aplicados efetivamente, abrem as portas para a estruturação semântica de bases de dados na web.

[2] VIAF - http://viaf.org/

[3] GeoNames - https://www.geonames.org/

[4] Painel de Informação do Google de Tim Berners-Lee - https://g.co/kgs/nb12Dg

Dessa forma, o papel do cientista da informação se revela imprescindível no contexto na web semântica, produzindo e apoiando estudos e projetos sobre "motores de busca, interfaces dos sistemas de informação, vocabulários controlados, indexação automática, gestão do conhecimento e inteligência competitiva" como elencado por Souza e Alvarenga (2004, p. 139-140).

E é, nessa perspectiva, que este artigo é situado, no contexto em que a Ciência da Informação apoia as instituições de patrimônio cultural na publicação de seus acervos digitais em rede, nesse caso em específico, propondo um estudo aplicado sobre o relacionamento dos dados dos acervos com provedores de dados semânticos.

E essa condição de relacionamento dos dados é posicionada sob o conceito de dados abertos ligados, que também é difundido por Berners-Lee (2006), e, em suma, é fundamentado sob quatro princípios: *Usar URIs como nomes para coisas; Usar URIs em HTTP para que as pessoas consigam acessar esses nomes; Prover informações úteis junto à URI, utilizando padrões semânticos, como RDF; Referenciar outras URIs, para que seja possível descobrir outras coisas.*

Esses princípios expressam de maneira objetiva e técnica que efetivar a publicação de dados abertos ligados envolve um processo de identificação dos objetos, como colocado acima, através de URIs (*Uniform Resource Identifier*) que são identificadores únicos na web, e certificar-se de que esses identificadores estejam acessíveis e disponíveis para serem ligados uns com os outros entre os provedores de dados da internet.

Essa estruturação de objetos na web é proveniente do contexto da web semântica, e, inclusive, sistematiza uma prática de sociabilidade em rede de informações entre as bases de conhecimento existentes, direcionando a produção e consumo de conteúdo na internet a uma maior versatilidade da relação entre o usuário e a máquina.

Dessa forma, imaginando promover essa sociabilidade em rede, algumas técnicas podem ser aplicadas para promover a contextualização semântica de bases de dados na web, uma delas é a reconciliação de dados, que, no contexto da web semântica, pode ser entendida como um dos processos inerentes ao enriquecimento semântico, que, por sua vez, constitui um conjunto de técnicas que objetivam ligar dados com bases de conhecimentos digitais (SANDERSON, 2016).

Essas bases de conhecimentos digitais podem ser compreendidas como sistemas de organização do conhecimento (KOS) disponíveis na web, e podem ser expressos tanto na forma de vocabulários controlados, quanto na forma de um repositório de objetos digitais estruturados semanticamente, como a Wikidata ou o *Geonames*, sendo chamados, nesse caso, de KOS-LOD, pois são sistemas de organização do conhecimento presentes na nuvem de dados abertos ligados (ZENG, 2019).

Uma boa forma de entender como ocorre o processo de enriquecimento semântico é refletir sobre o Framework de Enriquecimento Semântico proposto pela Europeana:

> *Análise*: a fase de pré-enriquecimento concentra-se na análise dos metadados originais, na seleção dos sistemas de conhecimento a serem ligados, e na proposição de regras para correspondência e vínculo dos metadados originais ao recurso textual disponíveis nos KOS selecionados; *Vinculação*: o processo de combinação automática entre os valores dos metadados com os valores dos objetos nos KOS, e a adição de relação contextuais entre os valores; *Acréscimo*: o processo de seleção dos valores do KOS a serem adicionados ao conjunto de dados original. Isso talvez não inclua somente conceitos em diferentes línguas, mas também conceitos mais específicos ou abrangentes (ISAAC *et al.*, 2015, p. 9).

Em síntese, o processo de enriquecimento semântico compreende um momento de análise dos metadados atuais do acervo/conjunto de dados a ser enriquecido, em que também é realizada a seleção de quais bases de conhecimento serão utilizadas como referência para a ligação dos dados e quais as especificações e regras para a efetivação da ligação dos dados.

Após a análise, o momento de vinculação se refere ao uso de técnicas automáticas/semiautomáticas de reconhecimento dos valores textuais provenientes dos metadados originais, nas bases de conhecimento on-line, bem como à descrição dos tipos de relacionamento das combinações resultantes das ligações. Por último, o acréscimo diz sobre a adição de informações presentes na base de conhecimento ao conjunto de dados enriquecido, com seus devidos relacionamentos expressos.

Esse processo que envolve o enriquecimento semântico, bem como a reconciliação de dados, é diretamente relacionado com a proposta dos dados abertos ligados, e é componente importante da estruturação e potencial aplicação em serviços futuros da web semântica entre as instituições de patrimônio cultural. Vale dizer que isso é ainda praticamente inexplorado nos repositórios digitais e serviços informacionais culturais no país, representando um ponto importante de desenvolvimento futuro para a área. É o processo que auxilia na mudança de concepção do modelo conceitual dos dados, saindo do foco no documento para o foco nas entidades. A partir da aplicação desse processo, o aprimoramento do acesso e reúso do acervo/conjunto de dados pode ser efetivado. Como consequência, pode-se obter novos serviços e produtos, gerando camadas de inovação a partir dos dados e ampliação do valor social dos acervos.

O estudo aplicado neste artigo tem o foco especificamente no tópico da reconciliação de dados, expressa na etapa de vinculação do enriquecimento semântico em que ocorre "o processo de combinação automática entre os valores dos metadados com os valores dos objetos nos KOS" (ISAAC *et al.*, 2015, p. 9), e, ainda no caso deste estudo, limita-se a implementar esse processo, sem, posteriormente, efetuar adição de relações contextuais.

O sistema de organização do conhecimento KOS-LOD utilizado neste experimento foi a Wikidata, parte do universo informacional da Wikipedia e que compartilha das mesmas características de colaboração em rede.

A Wikidata originalmente foi pensada para criar objetos digitais e relacioná-los ao conteúdo de páginas da Wikipedia , estruturada sob uma ótica semântica (<objeto><relação><objeto>). No entanto, a Wikidata logo ganhou independência e se tornou referência como base de conhecimento digital pela grande quantidade de informação estruturada e relacionada (VRANDEČIĆ; KRÖTZSCH, 2014). Atualmente, a Wikidata conta com mais de 71 milhões[5] de objetos digitais interligados em diversas linguagens.

METODOLOGIA

O método aplicado neste estudo se limita à forma de como foram concebidos os scripts de reconciliação de dados, já que estes foram desenvolvidos justamente para atender à expectativa do estudo de permitir entender melhor o processo de reconciliação semiautomática de dados, utilizando a linguagem de programação Python.

A operação de reconciliação de dados foi composta por dois scripts, sendo o primeiro deles, identificação de instâncias[6], com a função de identificar instâncias (classes de objetos) da Wikidata que melhor representam os valores de cada metadado da base de origem, uma vez sinalizada qual a instância. O segundo script de reconciliação de dados[7] realiza o processo de busca por objetos da Wikidata que representem os valores procurados, filtrando cada valor pela sua respectiva instância sinalizada no primeiro script, perfazendo, assim, um caminho de ligação de dados entre uma base de dados e um sistema de organização do conhecimento.

[5] Wikidata Statistics - https://www.Wikidata.org/wiki/Wikidata:Statistics/pt-br

[6] Script de identificação de instâncias - https://github.com/luisfeliperd/Wikidata_sparql/blob/master/Wikidata_metadata.py

[7] Script de reconciliação de dados – https://github.com/luisfeliperd/Wikidata_sparql/blob/master/Wikidata_value.py

O ambiente computacional utilizado para desenvolver os scripts foi constituído por um notebook com acesso à internet banda larga cabeada de 35Mb, cujo hardware é composto por um processador de 4 núcleos físicos de até 3.80GHz de frequência, 16GB de memória RAM, utilizando como unidade de armazenamento um SSD e como sistema operacional o Windows 10. Ressaltando ainda que não foi utilizado nenhum servidor de banco de dados, os dados coletados foram armazenados no formato CSV e processados posteriormente para análise em planilhas.

A título de experimentação, os dados utilizados nos scripts foram obtidos exportando a coleção de acervo museológico do Museu Histórico Nacional no formato CSV[8]. Essa base de dados é formada por 774 itens, e foram escolhidos três metadados para o processo de reconciliação, são eles: autor, técnica e material. A tabela 1 abaixo apresenta a quantidade de valores para cada metadado, sendo que essa quantidade não expressa o total de itens da base devido à existência de valores vazios, além disso, como os valores se repetem, a coluna *valores distintos* apresenta a quantidade de valores únicos sem repetição.

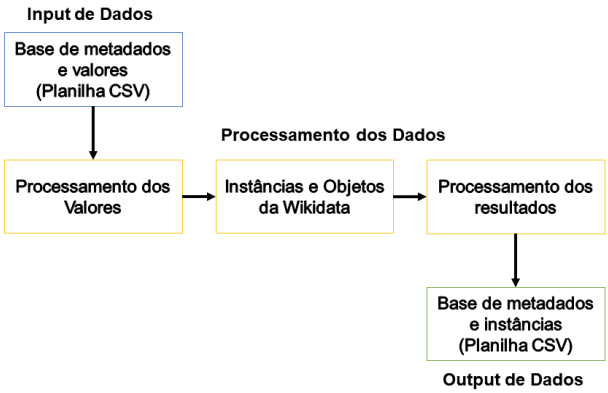
Tabela 1 – Síntese da quantidade de valores por metadado. Valores sintetizados a partir dos dados do acervo museológico do Museu Histórico Nacional

Metadado	Valores	Valores Distintos
Autor	646	241
Técnica	710	43
Material	775	51

Fonte: Dados da pesquisa (2020).

De forma geral, ambos os scripts funcionam a partir de um fluxo semelhante de processamento dos dados, diferenciado apenas nas funcionalidades aplicadas, como apresenta o fluxograma da figura 1.

Figura 1 – Fluxo básico dos scripts de reconciliação de dados



Fonte: Dados da pesquisa (2020).

O *input de dados* dos scripts ocorre através de uma planilha no formato CSV com a base de dados a ser reconciliada, no caso, no script de reconciliação de dados, a planilha de seleção de instâncias (resultante do script de identificação de instâncias) também é utilizada como entrada para definir em qual instância buscar de cada conjunto de valores.

O *processamento de dados* dos scripts perfaz o caminho de procurar por cada valor da base de dados de origem (acervo do MHN no caso) na Wikidata através de uma consulta SPARQL[9] no *endpoint* da API de consultas[10], que busca pelos valores nos rótulos de objetos na Wikidata. Essas consultas foram formadas, utilizando o tutorial produzido pela própria Wikidata[11], e o link de consulta SPARQL[12] para elaborar testes. No caso do script de identificação de instâncias, os resultados são focados nas possibilidades de identificar que tipo de instância seria sugerido para cada valor e, no caso do script de reconciliação de dados, os resultados obtidos de cada instância que expressam as possibilidades de objetos da Wikidata que referenciem o valor procurado.

[8] Os dados do acervo digital do museu podem ser acessados na função "Ver como..." da interface do Tainacan do MHN, por esse ponto de acesso, é possível acessar a base via API, uma planilha HTML e uma planilha CSV.

[9] O que são consultas SPARQL? - https://www.w3.org/TR/rdf-sparql-query/

[10] *Endpoint* da API de consulta SPARQL - https://query.Wikidata.org/sparql

[11] Tutorial SPARQL Wikidata - https://www.Wikidata.org/wiki/Wikidata:SPARQL_tutorial

[12] Link para consultas SPARQL na Wikidata - https://query.Wikidata.org/

Vale reforçar que instância aqui é considerada como um conceito abstrato ou uma entidade, por exemplo, Autor, e que o resultado de cada instância é, por exemplo, um autor específico ou um caso de uma instância, como Machado de Assis. Estruturar a busca por reconciliação dessa maneira amplia a possibilidade de filtro de uma base semântica, dado que primeiro se identifica a instância e depois se identificam os casos de uma instância que podem se conectar.

Já o *output de dados*, consiste no processamento dos resultados da consulta SPARQL, que são obtidos no formato JSON (Notação de Objetos JavaScript). No caso do script de identificação de instâncias, os dados de *nome da instância* e *QID da instância* (identificador dos objetos na Wikidata) são obtidos para cada valor procurado, e, para cada metadado, é mensurada a ocorrência das instâncias em seus valores, o que permite criar uma planilha com as instâncias que mais ocorreram para cada metadado, permitindo ao usuário escolher entre as indicações de qual instância melhor se adequa para representar os metadados da base de dados de origem. No caso do script de reconciliação de dados, cada valor da base de dados é buscado na Wikidata, de acordo com a instância selecionada para cada metadado na planilha resultante do script anterior, os termos dos valores são comparados com os termos dos rótulos dos objetos da Wikidata, e, de acordo com a semelhança entre os termos, uma pontuação é gerada, indicando qual objeto potencialmente representa o valor procurado. O método tem um princípio estatístico, descrito abaixo pela biblioteca *fuzzywuzzy*[13] , apresentando qual a maior probabilidade de um conjunto de objetos informacionais ser de uma determinada instância pela recorrência e proximidade de vários termos relacionados a ela, além de, dada uma instância, qual caso melhor se ajusta a um valor de ocorrência para um metadado específico, indicando o caso mais próximo de ocorrência.

A seguir, o desenvolvimento de cada script será descrito com base em 5 etapas: descrição das bibliotecas utilizadas, do processo de leitura dos dados da base de origem, da consulta dos valores na Wikidata e do processamento dos resultados obtidos.

Quanto ao script de identificação de instâncias, as bibliotecas utilizadas foram a *requests*[14] para fazer a consulta dos valores ao *endpoint* da API de consulta SPARQL da Wikidata, também foi utilizada a biblioteca *pandas*[15] para lidar com a estruturação dos dados no script, como leitura da base de dados, armazenamento dos dados coletados e exportação dos resultados, além dessas também foi utilizada a biblioteca *time*[16] para parar o script por alguns segundos a cada consulta à Wikidata, evitando um possível bloqueio de segurança da API, por fim, foi utilizada a biblioteca *datetime*[17] para calcular o tempo gasto para consultar os valores.

A leitura de dados para o script de identificação de instâncias ocorreu primeiramente lendo os registros da planilha da base de dados de experimentação do MHN, cuja composição foi mencionada na tabela 1, acima. Para cada metadado, foi recuperado cada valor e para cada valor realizada uma consulta. Não foi realizado nenhum processo de normalização dos dados, dessa forma, termos sem padronização, e valores em branco ocorrem nessa base de dados.

Outro ponto a ser ressaltado é que como cada objeto pode ter valores de metadados iguais, por exemplo, mesmos autores, ou mesma técnica de produção, os valores na base se repetem. Dessa forma, foi aplicada uma verificação no script que pula a consulta de um valor se ele já foi verificado anteriormente, isso reduz a demanda à API e otimiza a síntese dos resultados.

[13] Biblioteca *fuzzywuzzy* - https://pypi.org/project/fuzzywuzzy/

[14] Biblioteca *requests* - https://pypi.org/project/requests/#description

[15] Biblioteca *pandas* - https://pandas.pydata.org/docs/

[16] Biblitoeca *time* - https://docs.python.org/3/library/time.html

[17] Biblioteca datetime https://docs.python.org/3/library/datetime.html

Ainda foi necessário tratar a existência de múltiplos valores nos campos de técnica e material, por exemplo, alguns itens têm como material óleo e tela, o que ocorre na planilha com a notação "óleo||tela", com os termos divididos por *dual pipe* ("||"), desse modo, foi preciso definir um processo de separação desses valores e consulta a cada um individualmente.

Já na etapa de consulta dos valores na Wikidata, cada valor, após passar pelo processo de leitura de dados citado no parágrafo anterior, foi inserido em uma consulta SPARQL (figura 2) embutida no código. Essa consulta SPARQL foi estruturada para recuperar o identificador do objeto da Wikidata em "?sujeito", o identificador da instância do objeto encontrado em "?instancia_de_que" e o rótulo da instância em "?instancia_de_queLabel", a consulta é feita de forma unificada nos idiomas português de Portugal, português do Brasil e inglês, de forma que são consultados objetos da Wikidata que tenham o rótulo igual ao valor consultado. A consulta ainda prevê que não sejam recuperadas instâncias referentes à "categoria da Wikimedia" (Q4167836) e "páginas de desambiguação da Wikimedia" (Q4167410).

A cada consulta foi aplicada uma espera de 3 segundos visando a evitar problemas de bloqueio de segurança da API, além disso, o resultado da consulta pode retornar um JSON vazio se nenhuma instância for identificada, questão que foi tratada no script, pulando os valores que não retornaram resultados. Dessa forma, uma consulta válida retorna um JSON com dados sobre o nome da instância e o identificador da instância na Wikidata.

Figura 2 – Consulta SPARQL do script de identificação de instâncias

```
SELECT DISTINCT ?sujeito ?instancia_de_que ?instancia_de_queLabel WHERE {

                        { ?sujeito ?label "%s". }
                        UNION
                        { ?sujeito ?label "%s"@en. }
                        UNION
                        { ?sujeito ?label "%s"@pt-br. }

                        ?sujeito wdt:P31 ?instancia_de_que.

                        FILTER(?instancia_de_que NOT IN(wd:Q4167836, wd:Q4167410))

                        SERVICE wikibase:label { bd:serviceParam wikibase:language "pt-br", "pt", "en". }
}
```

Fonte: Dados da Pesquisa (2020).

Por fim, o processamento dos dados coletados foi executado em duas etapas, a primeira foi o armazenamento dos resultados de cada consulta SPARQL, juntamente com as informações dos valores consultados, o que consistiu em um dataframe (estrutura de dados de múltiplas variáveis no *pandas*) com o nome do metadado, o nome da instância e o identificador da instância recuperada. Após a coleta dos dados para todos os metadados, um novo dataframe foi construído, sintetizando os dados, foi calculada a ocorrência dos identificadores das instâncias, e consideradas apenas as 5 instâncias que mais se repetiram para cada metadado no esforço de identificar as instâncias mais representativas.

Uma planilha com essas 5 instâncias e suas respectivas ocorrências para cada metadado é o resultado deste script, essa mesma planilha contém uma coluna denominada "best_option" para que o usuário sinalize com um "x" qual das instâncias seria a mais representativa para cada metadado e, assim, utilizar essa planilha como *input* do próximo script, apontado em qual instância procurar os valores da base de dados para cada metadado.

Já quanto ao script de reconciliação de dados, foram utilizadas as mesmas bibliotecas *requests*, *pandas*, *time e datetime* utilizadas no script anterior, com a adição da biblioteca *collections* para usar especificamente a função *defaultdict*, que permite a criação de dicionários de listas para auxiliar na estruturação dos dados, e também a biblioteca *fuzzywuzzy* para calcular uma pontuação de semelhança de termos entre os valores da base de dados e os rótulos de objetos da Wikidata.

A leitura dos dados para esse script ocorreu da mesma forma como no script anterior, lendo a base de dados e verificando cada valor para cada metadado, atentando-se aos mesmos princípios de valores repetidos e a valores múltiplos. A diferença principal foi a leitura da planilha resultante do script de identificação de instâncias com a informação de qual instância usar para procurar os valores em cada metadado. A partir dela, o dado do identificador da instância foi adicionado à consulta SPARQL e permitiu pesquisar os valores em uma instância específica.

A etapa de consulta dos valores no script de reconciliação de dados consistiu na verificação dos valores de cada metadado da base de dados na Wikidata via consulta SPARQL (figura 3), que retornava como dados, o identificador do objeto na Wikidata ("?sujeito"), o rótulo principal do objeto ("?sujeitoLabel") e os rótulos alternativos do objeto ("?sujeitoAltLabel"). Essa consulta busca os valores da base de dados que tenham o termo semelhante aos rótulos dos objetos da Wikidata. Além disso, o dado da instância também é utilizado na penúltima linha da consulta para refinar a consulta dos valores a objetos da instância referente ao metadado.

Figura 3 – Consulta SPARQL do script de reconciliação de dados.

```
SELECT DISTINCT ?sujeito ?sujeitoLabel ?sujeitoAltLabel WHERE {

                                { ?sujeito ?label "%s". }
                                UNION
                                { ?sujeito ?label "%s"@en. }
                                UNION
                                { ?sujeito ?label "%s"@pt-br. }

                                ?sujeito wdt:P31 wd:%s

                                SERVICE wikibase:label { bd:serviceParam wikibase:language "pt-br", "pt", "en". }
}
```

Fonte: Dados da Pesquisa (2020).

Devido ao volume de consultas, foi preciso adicionar um *loop* ao script que verifica se a consulta à API foi bem sucedida, caso contrário é adicionada uma espera de 5 minutos até a realização da próxima consulta, visando a contornar um possível bloqueio momentâneo da API. Esse processo é repetido pelo menos 5 vezes, até que o acesso à API seja estabelecido ou, então, o script retorna ao erro de acesso.

O processamento dos resultados, no caso desse script, executa uma etapa de armazenamento dos identificadores dos objetos encontrados e dos seus respectivos rótulos, sendo que para cada rótulo um cálculo de semelhança entre o termo do valor consultado e os rótulos dos possíveis objetos da Wikidata encontrados é efetuado. Este cálculo é executado usando a biblioteca *fuzzywuzzy* que utiliza o princípio do cálculo da distância Levenshtein, cujo princípio "é definir a distância entre duas palavras com base no número de operações necessárias para torná-las iguais" (RUBERTO; RODRIGO, 2017, p. 31), essa operação no script retorna uma pontuação de semelhança entre termos no intervalo de 0 para nenhuma semelhança e 100 para termos iguais. Dessa forma, a segunda etapa do processamento de dados calcula uma média dos scores para cada objeto da Wikidata encontrado, apontando qual deles tem uma semelhança de termos maior com o valor consultado.

Por fim, o script resulta em uma planilha com os dados do nome do metadado, nome da instância referente a ele, o valor consultado, o identificador do objeto da Wikidata encontrado e a média da pontuação resultante do cálculo de semelhança entre os valores e os rótulos dos objetos. Com essa planilha, é possível analisar os resultados da reconciliação de dados com a Wikidata.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

Os resultados apresentados a seguir constituem a descrição dos produtos da consulta dos valores da base de dados do acervo do Museu Histórico Nacional (tabela 1) na Wikidata. Serão apresentados de forma sintética os resultados da identificação de instâncias e a descrição dos objetos da Wikidata identificados em relação aos valores da base de dados do MHN. A análise dos resultados tem o objetivo de complementar o entendimento do processo de reconciliação de dados através de programação em Python, observando como os produtos dos scripts completam o ciclo de ligação de dados.

Analisando o tempo gasto para a reconciliação dos dados (tabela 2), houve uma semelhança entre os períodos dos processamentos de consulta de dados. Observando que para o metadado "Autor" houve mais tempo gasto, porém também houveram mais itens consultados, mesma condição de proporção se repete para a consulta dos valores dos demais metadados. Analisando a coluna "Média por item", que calcula a divisão entre o tempo gasto e a quantidade de itens consultados, observa-se que, independente do script, uma consulta na Wikidata leva em média de 3 a 4 segundos para ser executada, levando em conta ainda que o script adiciona um tempo de espera de 3 segundos para evitar o bloqueio de segurança da API, se não fosse essa adição de tempo, a consulta levaria um segundo ou menos para ser executada.

Esse indicador mostra a que a aplicação do script é viável com tempo de resposta considerado baixo em bases com um volume relativamente pequeno de dados, levando em conta tanto a quantidade de valores distintos que se deseja reconciliar (até 5.000) se o tempo de execução mantiver a proporção de 3 segundos por consulta.

Para afirmar com mais certeza essa viabilidade, seria interessante investir em uma análise com diferentes quantidades e tipos de conjuntos de dados, para identificar se há alguma limitação por volume e sua viabilidade em conjuntos de dados com uma grande quantidade de observações (5.000 ou mais), em que tal serviço pode ser executado com diferentes estratégias, não impactando na experiência final de navegação do usuário e permitindo um enriquecimento contínuo de acervos mesmo de grande volume.

No entanto, é importante ressaltar que essa estimativa se aplica ao contexto da reconciliação de dados em massa. Uma vez que a reconciliação seja incorporada ao sistema de indexação de objetos, ou que a instituição adote essa etapa de enriquecimento como parte da indexação de objetos digitais na web, espera-se um volume de dados mais baixo e contínuo, abrindo espaço para que o usuário valide os resultados da reconciliação por exemplo.

Tabela 2 – Tempo de execução dos scripts de reconciliação de dados

Script	Metadado	Tempo de verificação	Média por item
Identificação de instâncias	Autor	00:14:30	00:00:04
Reconciliação de dados	Autor	00:14:28	00:00:04
Identificação de instâncias	Técnica	00:02:30	00:00:03
Reconciliação de dados	Técnica	00:02:21	00:00:03
Identificação de instâncias	Material	00:02:56	00:00:03
Reconciliação de dados	Material	00:03:01	00:00:04
TOTAL		00:39:46	-

Fonte: Dados da Pesquisa (2020).

Vale ainda deixar claro que este estudo não tinha o objetivo de mensurar os níveis de viabilidade da reconciliação de dados em grande escala e, por isso, não se têm resultados suficientes para declarar qual infraestrutura tecnológica melhor se adequaria a contextos com uma quantidade de dados maior.

Observando os resultados do script de identificação de instâncias (tabela 3), um volume baixo de instâncias foi identificado para os metadados "Material" e "Técnica", sendo que este último não apresentou nenhuma instância que ocorreu mais de uma vez, já o metadado "Autor" apresentou instâncias com ocorrências maiores, as ocorrências indicam inicialmente que houve mais facilidade de identificar objetos na Wikidata com rótulos semelhantes aos valores de "Autor" do que de "Material" e "Técnica".

Tabela 3 – Resultados do script de identificação de instâncias

Metadado	Instâncias Propostas	Ocorrências
Material	ser humano	4
	elemento químico	4
	táxon	2
	material	2
	fibra	2
Autor	ser humano	205
Metadado	Instâncias Propostas	Ocorrências
Autor	artigo científico	121
	sobrenome	10
	ortsteil	7
	obra criativa	7
Técnica	visual arts technique	1
	setor econômico	1
	Página web	1
	técnica artística	1
	atividade	1

Fonte: Dados da pesquisa (2020).

Outro ponto interessante é observado nas instâncias sugeridas, como no caso do metadado "Material", com 4 ocorrências da instância "ser humano", que, inicialmente, não tem relação com os valores desse metadado, já os demais possuem algum tipo de relação. Vale destacar que somente a instância "material", que foi a selecionada, aponta relação explícita, porém só ocorreu 2 vezes. Quanto ao metadado "Técnica", somente as instâncias "visual arts technique" e "técnica artística" (selecionada) têm relação efetiva com os valores do metadado, porém só ocorreram uma vez, o que já indica uma baixa conexão com os dados da Wikidata. Isso evidencia áreas da base de conhecimento que precisam ser potencialmente melhoradas e complementadas com novos conjuntos de dados. Por último, o metadado "Autor" apresenta inicialmente duas instâncias relacionadas com os valores do metadado, "ser humano" e "sobrenome", sendo o primeiro com maior incidência, foi o escolhido para representar o metadado.

Esses resultados já indicam um nível baixo de reconhecimento dos valores dos metadados "Material" e "Técnica" na Wikidata e um nível mais alto para o metadado "Autor", o que já aponta uma maior recuperação de objetos que se relacionem com os valores de autor que os demais metadados. Isso sugere o questionamento para identificar a causa desses resultados, se a Wikidata realmente tem uma lacuna de objetos na língua portuguesa sobre esses tipos de metadados, ou se existe um arranjo melhor para identificar os objetos. Explorar bases de conhecimento dessa maneira pode gerar panoramas de áreas de conhecimento nos quais uma base pode ser melhor aplicada em relação a outras, permitindo avançar o conhecimento em seu potencial de uso.

Já observando os resultados do script de reconciliação de objetos, as indicações constatadas acima são confirmadas, como apresenta a tabela 4. O metadado "Autor" apresentou uma maior proporção de objetos identificados, diferente da proporção de identificação dos metadados "Material" e "Técnica", com um resultado consideravelmente menor.

Tabela 4 – Resultados do script de reconciliação de objetos

Metadado	Nº de Itens Distintos	Nº de Itens com Objetos Identificados	Proporção
Autor	243	157	64,61%
Material	51	2	3,92%
Técnica	43	1	2,33%

Fonte: Dados da pesquisa (2020).

Este estudo não produziu dados suficiente para identificar o motivo da baixa proporção de objetos da Wikidata identificados para os metadados "Material" e "Técnica", o que abre a premissa para a investigação dessa condição em novos estudos, cujo objetivo seja verificar, por exemplo, quais motivos levaram a este resultado.

Em busca de verificar a qualidade dos valores ligados a objetos da Wikidata, uma validação dos objetos foi realizada, identificando quais objetos realmente tinham relação com os valores consultados. Para o metadado "Técnica", o único objeto identificado foi relativo ao valor "Pintura a óleo", identificado na Wikidata como o "oil painting" de identificador "Q56676227", constituindo uma relação válida. Já no metadado "Material", os dois valores com objetos identificados na Wikidata efetivaram relações válidas: o valor "vidro" com o objeto "glass" de identificador "Q11469" e o valor "alumínio" com o objeto "aluminum" de identificador "Q663".

No metadado "Autor", dos 157 valores identificados, 143 (91,08%) deles tiveram a relação válida com os objetos da Wikidata, os demais 14 não obtiveram resultados válidos. A validade no caso deste metadado foi verificada analisando qualitativamente o objeto e identificando se o mesmo contém informações que indicam de maneira positiva a referência ao valor consultado, por exemplo, pessoas com "ocupações" relacionadas à cultura, como pintores, ilustradores ou compositores. É importante ressaltar que a taxa de acerto apresenta um resultado bastante significativo para essa instância, mostrando um potencial a ser explorado para serviços informacionais. Outro ponto interessante a ser observado é a relação entre a pontuação de semelhança dos termos do valor consultado com o rótulo do objeto da Wikidata e a validade dos objetos. O objetivo de se calcular a semelhança entre os termos foi de indicar qual objeto tem a maior possibilidade de manter uma relação válida com o valor consultado, mas no caso dos valores do metadado "Autor", a correlação entre essas duas variáveis foi muito baixa, de 0,30[18].

Uma nova pesquisa pode investigar melhor as condições de aproximação entre os valores consultados e a validade das relações com os objetos da Wikidata, se a aproximação textual é o caminho mais viável e quais são as outras possibilidades. É importante ressaltar que identificar heurísticas mais adequadas para validar esses resultados de forma automática ou semiautomática é um objetivo de pesquisa a ser explorado futuramente.

Ainda, algumas observações do processo de validação são interessantes de serem elencadas, como a importância de um preenchimento consistente dos valores. Em muitos casos, o valor do nome do autor continha somente um sobrenome ou um só nome e não o nome completo, o que dificulta a busca. Por exemplo, no caso do autor de nome "Otto" 14 objetos da Wikidata foram indicados, mas nenhum foi validado, já para os nomes completos, a ocorrência de objetos únicos e válidos foi maior.

Outra observação relevante é a capacidade do processo da Wikidata de desambiguação de valores, uma vez que a base não foi previamente normalizada, alguns nomes iguais, mas digitados de maneira diferente foram identificados como o mesmo objeto na Wikidata, o que, no final das contas, desambigua a base por si. Por exemplo, os nomes "Spanyi Ernest César Novak" e "Spanyi Ernesto César Agostinho Novak" são o mesmo autor, porém escritos de forma diferente na base de dados, ao reconhecer o mesmo objeto da Wikidata para os dois valores, este autor é desambiguado. Essa questão suscita a discussão sobre a importância da normalização dos dados e da aplicação de regras de catalogação consistentes em um processo de abertura e ligação de dados.

Um último ponto interessante é a capacidade importante que a reconciliação de dados com a Wikidata tem de reconhecer os valores consultados. Como cada objeto na Wikidata tem um rótulo e a possibilidade de rótulos alternativos, um objeto pode ser reconhecido através de suas variações terminológicas. Assim, objetos com nomes diferentes em determinadas ocasiões podem ser reconciliados e identificados como o mesmo objeto.

[18] Correlação de Pearson, calculada entre a matriz de pontuação da semelhança de termos e a matriz de validação dos objetos da Wikidata.

Por exemplo, o nome de autor "Serrano", cujo rótulo do objeto na Wikidata é "Giovanni Battista Crespi", porém pode ser reconhecido nos rótulos alternativos como "Serrano". Essa condição permite a ligação de dados entre diferentes provedores, sem perder a identidade do valor na base de origem e conservando uma única identidade ao objeto.

CONCLUSÕES

Ao refletir sobre o processo de desenvolvimento dos scripts de reconciliação de dados e seus resultados, alguns tópicos importantes demandam mais atenção:

- É nítido que um tratamento prévio dos dados e o cuidado com as regras de catalogação podem aumentar o potencial da reconciliação de dados. Como apontado nos resultados acima, algumas características indicavam a falta de normalização da base de dados, como valores vazios, valores que são iguais escritos de maneira diferentes, e, no caso dos autores, nomes incompletos. Essas condições dificultam a busca pelos termos, quanto mais completo e menos ambíguo um valor é, mais fácil será de encontrar um objeto digital correspondente em um sistema de organização do conhecimento como a Wikidata.

- Uma questão que paira ao observar a proporção de objetos efetivamente reconciliados dos metadados de "Material" e "Técnica" é o quanto a Wikidata pode ser entendida como referência para efetivar a ligação de dados com objetos digitais culturais na língua portuguesa? De acordo com os resultados, é clara a relevância da Wikidata ao buscar pelos autores, mas quanto aos outros metadados, seria necessária uma análise mais profunda das causas do baixo retorno obtido, e que confirme quais alternativas ou outras formas de abordagem podem ser elaboradas para reconciliar esses tipos de dados.

- Outro ponto de relevância indiscutível são os benefícios de se reconciliar os dados com sistemas digitais de organização do conhecimento como a Wikidata. O primeiro motivo é o auxílio à normalização dos dados com a desambiguação de termos, como cada objeto na Wikidata tem uma lista de rótulos alternativos, as diferentes variações de escrita são atendidas e, se por acaso um mesmo item aparece escrito de duas maneiras diferentes na base de dados de origem, eles podem estar contemplados no conjunto de rótulos alternativos, e serão entendidos como um único objeto digital. Outro benefício importante é o enriquecimento de dados, uma vez que os dados estão ligados, informações que não foram contempladas na base de dados de origem podem ser recuperadas da Wikidata, auxiliando a contextualizar melhor os objetos culturais, permitindo filtros por categorias antes impensáveis, como por exemplo, itens por nacionalidade do autor, ou itens cujos autores são escritores e, ainda, viabilizando a conexão com outros acervos culturais através do indicador de identificadores externos, que apontam a existência de determinado objeto em outros provedores de dados.

Entende-se, então, que o processo de reconciliação de dados, envolve a qualidade dos dados de origem, a conexão com um sistema digital de organização de conhecimento que permite consultas, formas de elencar quais resultados são mais relevantes, como a frequência das instâncias e a pontuação por semelhança entre termos e uma validação dos resultados obtidos, para, efetivamente, realizar a ligação dos dados. Cada etapa deste processo pode derivar um aprofundamento específico em busca de melhorar a forma como é utilizada e auxiliar na efetivação do processo como um todo, em busca de promover a socialização da informação que uma vez existe isolada em repositórios e portais institucionais. Promover a reconciliação dos dados, então, além de enriquecer o acervo de origem e promover a ideia dos dados abertos ligados, proporciona à sociedade seu direito de acesso à informação e, além disso, à suas raízes culturais.

Usar o meio web para promover esse tipo de fenômeno é dar ao público o que o pertence, de maneira contextualizada e interconectada, favorecendo o compartilhamento de conhecimento e desenvolvendo a interação social através da cultura digital em rede.

Dessa forma, entende-se que o processo de desenvolvimento e descrição dos scripts de reconciliação de dados atende ao objetivo proposto inicialmente, o de contribuir para o entendimento de como se dá a aplicação de uma técnica de enriquecimento semântico em bases de dados culturais, e responde ao anseio de entender como se dá o processo de reconciliação de dados culturais com a Wikidata através da programação de scripts na linguagem Python. Ao fim do desenvolvimento e experimentação dos scripts, conclui-se que eles podem ser aplicados futuramente em outros contexto de reconciliação de conjuntos de dados, em busca de fundamentar a ligação e a abertura dos dados, bem como mediar o enriquecimento da base de dados de origem, uma vez que dada a efetivação da ligação de dados, dados do sistema de organização do conhecimento (Wikidata) podem ser adicionados ao conjunto de origem, e o processo reverso também pode ocorrer, dados do conjunto de origem podem ser inseridos na Wikidata se já não existirem.

REFERÊNCIAS

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. *Scientific American*, v. 284, n. 5, p. 34-43, 2001.

BERNERS-LEE, T. *Linked data principles,* 2006. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html> Acesso em: 10 set. 2020.

ISAAC, A.; MANGUINHAS, H.; STILLER, J.; CHARLES, V. *Report on enrichment and evaluation.* The Hague, Netherlands: Europeana Task Force on Enrichment and Evaluation, 2015. Disponível em: <http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Enrichment_Evaluation/FinalReport_EnrichmentEvaluation_102015.pdf>. Acesso em 15 de abr. de 2020.

RUBERTO, D.L.V.G.; ANTONIAZZI, R. L. Análise e Comparação de Algoritmos de Similaridade e Distância entre strings Adaptados ao Português Brasileiro. In: ANAIS DA XIII ESCOLA REGIONAL DE BANCO DE DADOS, XIII, 2017, [s.l]. *Anais…* [s.l]: SBC, 2017.

SANDERSON, R. *"The Linked Data Snowball and Why We Need Reconciliation"*, 2016. Disponível em: <https://www.slideshare.net/azaroth42/linked-data-snowball-or-why-we-need-reconciliation.> Acesso em: 28 mar. 2020.

SANTARÉM SEGUNDO, J. E. Web Semântica: Introdução a recuperação de dados usando SPARQL. In: *Encontro Nacional de Pesquisas em Ciência da Informação (ENANCIB),* v. 14, p. 3242-3261, 2014.

SOUZA, R. R.; ALVARENGA, L. A Web Semântica e suas contribuições para a ciência da informação. *Ciência da Informação*, v. 33, n. 1, p. 132-141, 2004.

VRANDEČIĆ, D.; KRÖTZSCH, M. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, v. 57, n. 10, p. 78-85, 2014.

ZENG, M. L. Semantic enrichment for enhancing LAM data and supporting digital humanities. Review article. *El profesional de la información*, v. 28, n. 1, 2019.

AGRADECIMENTOS

Agradecimentos à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) por fomentar a produção deste artigo através do programa de bolsa de pesquisa para doutorandos, e ao grupo de pesquisa Laboratório de Inteligência de Redes (UnB) por fomentar e conceder infraestrutura para a aplicação desta pesquisa

Recuperação de informação: descoberta e análise de *workflows* para agregação de dados do patrimônio cultural

Joyce Siqueira
Doutoranda em Ciência da Informação pela Universidade de Brasília (UnB) - Brasília, DF – Brasil. Mestre em Ciência da Computação pela Universidade Federal de Goiás (UFG) – GO - Brasil. Professora da Universidade Católica de Brasília (UCB) - Brasília, DF – Brasil.
http://lattes.cnpq.br/7006325340635761
E-mail: joycitta@gmail.com

Dalton Lopes Martins
Pós-Doutorado pela Universidade de São Paulo (USP) – SP - Brasil. Doutor em Ciência da Informação pela Universidade de São Paulo (USP) - São Paulo, SP - Brasil. Professor da Universidade de Brasília (UnB) - Brasília, DF - Brasil.
http://lattes.cnpq.br/3774617443225038
E-mail: dmartins@gmail.com

Submetido em: 28/10/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021.

RESUMO

Nos últimos anos, diferentes instituições culturais vêm envidando esforços para difundir a cultura por meio da construção de uma interface única de busca, que integre objetos digitais e facilite a recuperação de dados para os usuários leigos. Contudo, integrar dados culturais não é uma tarefa trivial, pois estes são diversos e singulares, necessitando de uma variedade de etapas entre a coleta e a apresentação. Com objetivo de identificar estas etapas, esta pesquisa visa localizar *workflows* de agregação de dados e discuti-los. Para tal, realizou-se pesquisa descritiva e bibliográfica, de natureza qualitativa, em bases de dados acadêmicas e na literatura cinzenta. Como resultado, apresentam-se oito projetos: American Art Collaborative, DigitalNZ, D-NET Software, Europeana, Mexicana, Parthenos Aggregator, TROVE e UNLV's Linked Data Project. A análise do conjunto de *workflows* resultou em oito diferentes etapas a serem executadas: 1. Extrair, 2. Estruturar, 3. Transformar, 4. Reconciliar, 5. Armazenar, 6. Publicar, 7. Expor e 8. Possibilitar novas aplicações. Além disso, também é visível a necessidade de maior detalhamento das etapas, a fim de que seja possível replicar o *workflow*, e usufruir de seus benefícios em outras instituições.

Palavras-chave: Agregação de dados. Busca integrada. Patrimônio cultural. Recuperação de informação. *Workflow*.

Information retrieval: discovery and analysis of workflows for aggregating cultural heritage data

ABSTRACT

In recent years, different cultural institutions have been making efforts to spread culture through the construction of a unique search interface, which integrates digital objects and facilitates data retrieval for lay users. However, integrating cultural data is not a trivial task, as these are diverse and unique, requiring a variety of steps between collection and presentation. In order to identify these steps, this research aims to locate data aggregation workflows and discuss them. To this end, descriptive and bibliographic research, of a qualitative nature, was carried out in academic databases and in gray literature. As a result, eight projects are presented: American Art Collaborative, DigitalNZ, D-NET Software, Europeana, Mexicana, Parthenos Aggregator, TROVE and UNLV's Linked Data Project. The analysis of the set of workflows resulted in eight different steps to be performed: 1. Extract, 2. Structure, 3. Transform, 4. Reconcile, 5. Store, 6. Publish, 7. Expose and 8. Enable new applications. In addition, the need for more detailed stages is also visible, so that it is possible to replicate the workflow, and enjoy its benefits in other institutions..

Keywords*: Data aggregation. Integrated search. Cultural heritage. Information retrieval. Workflow.*

Recuperación de información: descubrimiento y análisis de flujos de trabajo para agregar datos del patrimonio cultural

RESUMEN

En los últimos años, diferentes instituciones culturales se han esforzado por difundir la cultura mediante la construcción de una interfaz de búsqueda única, que integra objetos digitales y facilita la recuperación de datos para usuarios legos. Sin embargo, la integración de datos culturales no es una tarea trivial, ya que son diversos y únicos, y requieren una variedad de pasos entre la recopilación y la presentación. Para identificar estos pasos, esta investigación tiene como objetivo localizar los flujos de trabajo de agregación de datos y discutirlos. Para ello, se realizó una investigación descriptiva y bibliográfica, de carácter cualitativo, en bases de datos académicas y en literatura gris. Como resultado, se presentan ocho proyectos: American Art Collaborative, DigitalNZ, D-NET Software, Europeana, Mexicana, Parthenos Aggregator, TROVE y Linked Data Project de UNLV. El análisis del conjunto de flujos de trabajo resultó en ocho pasos diferentes a realizar: 1. Extraer, 2. Estructurar, 3. Transformar, 4. Reconciliar, 5. Almacenar, 6. Publicar, 7. Exponer y 8. Habilitar nuevas aplicaciones. Además, también es visible la necesidad de etapas más detalladas, para que sea posible replicar el flujo de trabajo y disfrutar de sus beneficios en otras instituciones.

Palabras clave: *Agregación de datos. Búsqueda integrada. Patrimonio cultural. Recuperación de información. Flujo de trabajo.*

INTRODUÇÃO

Instituições culturais estão, a cada dia, reinventando-se e inovando suas formas de interagir com público, com destaque, a disponibilização de objetos digitais e seus metadados em sites e/ou repositórios institucionais, como um meio para exercer sua prática comunicacional e difundir seus acervos digitalizados.

Essa realidade fez explodir, no Brasil e no mundo, a quantidade de objetos na rede, resultando em uma nova problemática: como permitir que os usuários, principalmente os leigos, encontrem o objeto de seu interesse, em meio a tanta oferta e a diferentes mecanismos de busca?

De forma ampla, a resposta a esta pergunta foi oferecer uma interface de busca integrada, que agrega um conjunto específico de bancos de dados, capaz de recuperar, mais facilmente, o objeto desejado. Com esse intento, nos anos 2000, algumas bibliotecas adotaram a pesquisa federada, que realiza a busca simultânea em diversas fontes, apresentando os resultados em uma lista única. No entanto, com o tempo, tornou-se evidente uma série de problemas, tais como: lentidão nos tempos de resposta; resultados duplicados e a impossibilidade de refinamento dos resultados (BRIGHAM *et al.*, 2016; PAVÃO; CAREGNATO, 2015).

Dessa forma, difundir a cultura por meio da oferta de uma interface de busca integrada, com uma navegação eficiente ainda é um objetivo fortemente almejado e, falando especificamente de Brasil, algo que ainda não foi realizado em ampla escala e que poderia contribuir, de forma significativa, para outras formas de socialização da cultura brasileira.

A agregação de dados culturais não é uma tarefa trivial, pois os metadados e objetos digitais são diversos e singulares, dificultando, sobremaneira, a definição de padrões. Apesar de diversos padrões de metadados, modelos conceituais e regras de catalogação, tais como CIDOC-CRM, EDM, LRM, entre outros, existirem para a área da cultura, os mesmos nem sempre são consensuados e se encontram aplicados em níveis muito diferentes de interiorização pelas instituições. Cabe ressaltar, a título de explicitação, que se considera neste trabalho que agregação de dados envolve a agregação de metadados mais a agregação dos objetos digitais descritos por esses metadados.

A Europa, por exemplo, lançou, em 2008, o protótipo Europeana, que deu acesso, logo no lançamento, a 4.5 milhões de objetos digitais de bibliotecas, museus, arquivos audiovisuais e galerias. Em 2020, fornece acesso a 58 milhões de objetos digitais, com sofisticadas ferramentas de pesquisa e filtro, além de coleções temáticas, exposições, galerias e blogs (EUROPEANA, 2020, *on-line*). A Europeana é um caso mundialmente conhecido, faz parte dos resultados deste estudo, contudo, outras instituições também realizam pesquisa na área e oferecem soluções para agregação de dados, considerando diferentes realidades.

Assim, o objetivo deste estudo é localizar e discutir *workflows* de agregação de dados culturais, para realizar uma análise qualitativa das etapas escolhidas por cada instituição, por meio de pesquisa de caráter descritivo e bibliográfico, de natureza qualitativa, realizada em bases de dados acadêmicas e na literatura cinzenta.

Para melhor compreensão do objetivo da pesquisa, *workflow* pode ser definido como:

> uma coleção de atividades organizadas para realizar um processo, quase sempre de negócio. Essas atividades podem ser executadas por um ou mais sistemas de computador, por um ou mais agentes humanos ou de software, ou então por uma combinação destes. Do que consistem, a ordem de execução e as pré-condições das atividades estão definidas no workflow, sendo que o mesmo é capaz ainda de representar a sincronização das atividades e o fluxo de informações entre elas (PEREIRA e CASANOVA, 2003, p. 1).

Ao final, foram localizados oito *workflows* que são apresentados na seção de Resultados, assim como a descrição das etapas. Este artigo está assim dividido: seção 2, Metodologia, seção 3, Análise e Discussão dos Resultados, e por último, na seção 4, as Conclusões.

METODOLOGIA

Pesquisa de caráter descritivo e bibliográfico, de natureza qualitativa, realizada em bases de dados acadêmicas e na literatura cinzenta, com o intuito de encontrar *workflows* de agregação de dados culturais.

As buscas foram realizadas no Google, EBSCOhost e BRAPCI, utilizando as palavras: *"pipeline", "architecture", "aggregation", "metadata ingest", "metadata aggregation", "aggregative data infrastructures"* e suas versões em português. Além destes, também foram realizadas pesquisas por meio de projetos de agregação de dados conhecidos, como: *"Europeana", "Mexicana", "Digital Public Library of America", "Trove" e "DigitalNZ"*. Cabe dizer que estes projetos foram escolhidos por serem os principais agregadores de dados culturais nas diferentes regiões do mundo, conforme apresentado por Navarrete (2016).

Optou-se pelo Google para localizar *workflows* na literatura cinzenta, a BRAPCI, por ser específica da área de Ciência da Informação no Brasil e a EBSCOhost, pela produção científica internacional, tornando a pesquisa mais ampla.

DESCRIÇÃO DOS WORKFLOWS

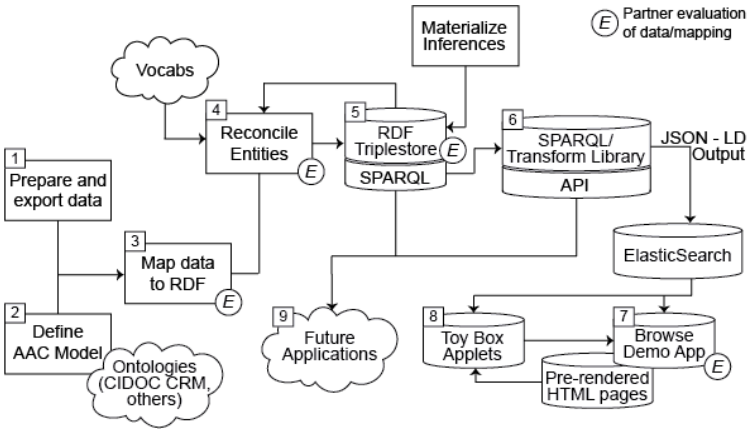
As pesquisas resultaram em oito projetos de agregação, realizados por sete instituições, listadas no quadro 1, cujos *workflows* são apresentados e detalhados nesta seção. Em alguns projetos, não foram encontrados os *workflows*, mas sim sua arquitetura, que, por mostrar etapas, foram considerados neste estudo.

Quadro 1 − Instituições, países de origem e seus projetos

N.	Instituição	País	Projeto
01	American Art Collaborative	EUA	AAC
02	Biblioteca Nacional	Austrália	Trove
03	Biblioteca Nacional	Nova Zelândia	DigitalNZ
04	Fundação Europeana	União Europeia	Europeana
05	Instituto de Ciência e Tecnologia da Informação	Itália	D-NET Software
06	Instituto de Ciência e Tecnologia da Informação	Itália	Parthenos Aggregator
07	Secretaria de Cultura	México	Repositório Mexicana
08	Universidade de Nevada	EUA	UNLV's Linked Data Project

Fonte: Elaborado pelos autores (2020).

Figura 1 − Workflow de agregação da AAC



Fonte: Fink (2018, p. 32). Adaptada.

AMERICAN ART COLLABORATIVE - AAC PIPELINE

A *American Art Collaborative* (AAC) é um consórcio de 14 instituições de arte que visam a investigar e a começar a construir uma massa crítica de *Linked Open Data* (LOD). Para Fink (2018), LOD se trata de um método para publicar dados estruturados na web de forma que as informações sejam interconectadas e, assim, tornadas amplamente úteis. A figura 1, apresenta o *workflow* proposto pela AAC.

O *workflow* prevê nove etapas. A Etapa 1."*Prepare and export data*", em tradução livre, "Preparar e exportar os dados", visa a fornecer dados principais e dados adicionais, considerados úteis pelos parceiros, que, na prática, exportam dados brutos de seus sistemas de origem e os carregaram em um repositório GitHub compartilhado. A etapa 2. "*Define AAC Model*" ou "Definir o Modelo AAC", trata-se de um conjunto geral de orientações sobre ontologias que podem ser adotadas e reutilizadas, tendo por objetivo constituir um modelo conceitual único para agregar os dados das diferentes instituições.

A etapa 3. "*Map data to RDF*" ou "Mapear dos dados para RDF", visa a mapear os dados das instituições parceiras para um modelo de destino e a fornecer flexibilidade para usuários adicionais, assim, usando o modelo de destino e a ferramenta de integração de dados Karma, os dados de cada parceiro são convertidos em RDF. A etapa 4. "*Reconcilie Entities*" ou "Reconciliar entidades" visa a mapear entidades individuais para IDs comuns, considerando, sempre que possível, vocabulários públicos padronizados. A etapa 5. "*RDF Triple Store - SPARQL*" ou "Armazenar dados RDF em um Triple Store - SPARQL", visa a armazenar e a fornecer acesso aos dados vinculados enriquecidos pela reconciliação de entidades e permite consultas SPARQL. A etapa 6. "*SPARQL/Transform library - API*" ou "SPARQL/API para transformação de Bibliotecas", visa a compilar um conjunto de consultas claras, reutilizáveis e focadas em entidades que navegam no gráfico de dados vinculados para fornecer documentos JSON-LD para desenvolvedores e humanistas digitais.

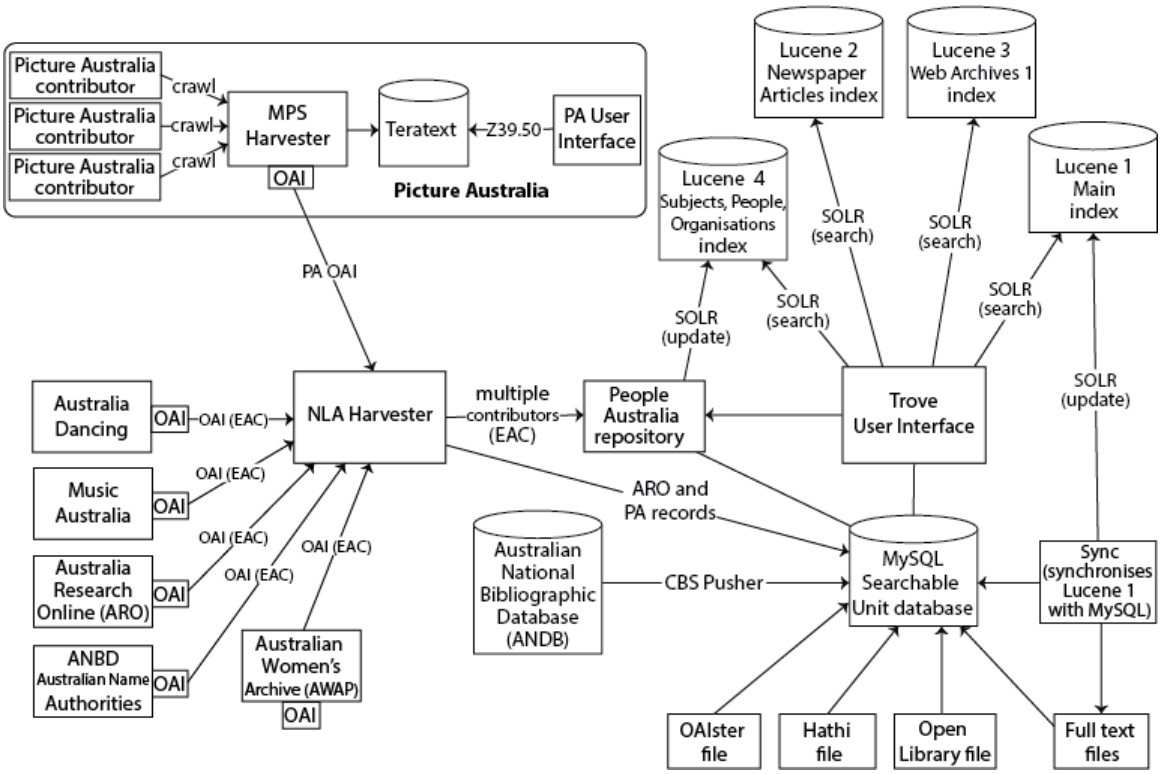
A etapa 7. "*Browse Demo App*" ou "Aplicações via browsers", visa a apresentar uma ilustração inicial suficientemente rica de dados vinculados para os usuários. A etapa 8. "*Toy Box Applets*" ou "Ampliação das aplicações via browser", visa a ampliar as possibilidades e ideias do AAC, por meio de um aplicativo de navegação. A etapa 9. "*Future Applications*" ou "Produzir aplicações futuras" visa a continuar a explorar casos de uso e aplicativos para dados vinculados por meio de contribuições de dados de parceiros (FINK, 2018).

BIBLIOTECA NACIONAL DA AUSTRÁLIA – TROVE

A Biblioteca Nacional da Austrália desenvolveu o Trove, que objetiva disponibilizar recursos culturais relacionados à Austrália. O Trove oferece um mecanismo de busca integrada em bibliotecas, museus, arquivos e outras organizações de pesquisa, além de um conjunto de serviços (TROVE HELP CENTRE, 2020). A figura 2 apresenta o *workflow* proposto.

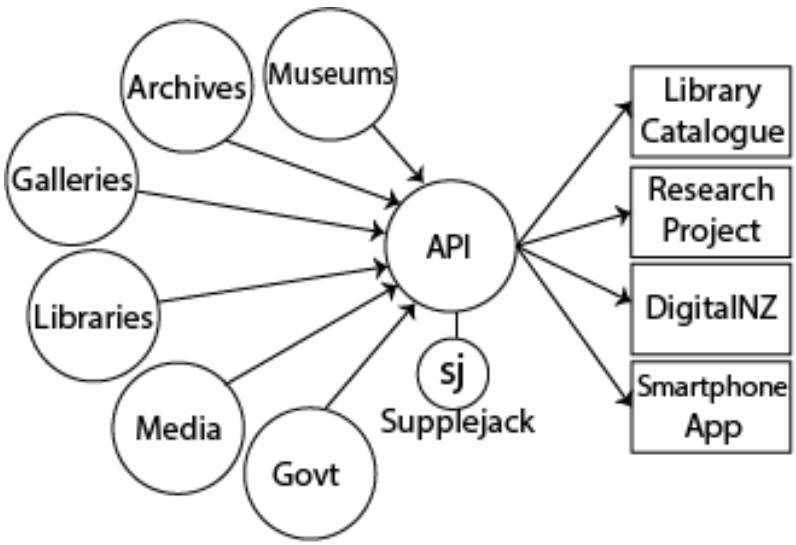
O Trove mantém um site com informações relevantes do projeto, no entanto, ainda que apresente o *workflow* de agregação, não traz, nas fontes pesquisadas, documentação que explique cada etapa do *workflow*. Contudo, na análise realizada pela pesquisa, percebe-se um grande uso do protocolo OAI-PMH para coleta de dados dos provedores, bem como a criação de diferentes índices para indexação ágil e recuperação da informação utilizando a tecnologia Apache Lucene, que é um software voltado para busca e indexação de documentos de alta escalabilidade e aplicado em projetos que exigem processamento de dados massivos. A arquitetura se vale de diversas camadas, mas devido à falta de documentação explícita, somente se pode inferir como as camadas se relacionam, sem condições de uma análise crítica da solução para eventual replicação.

Figura 2 – Workflow de agregação da Trove



Fonte: National Library of Australia (2010). Adaptada.

Figura 3 – Workflow de agregação proposto pela DigitalNZ



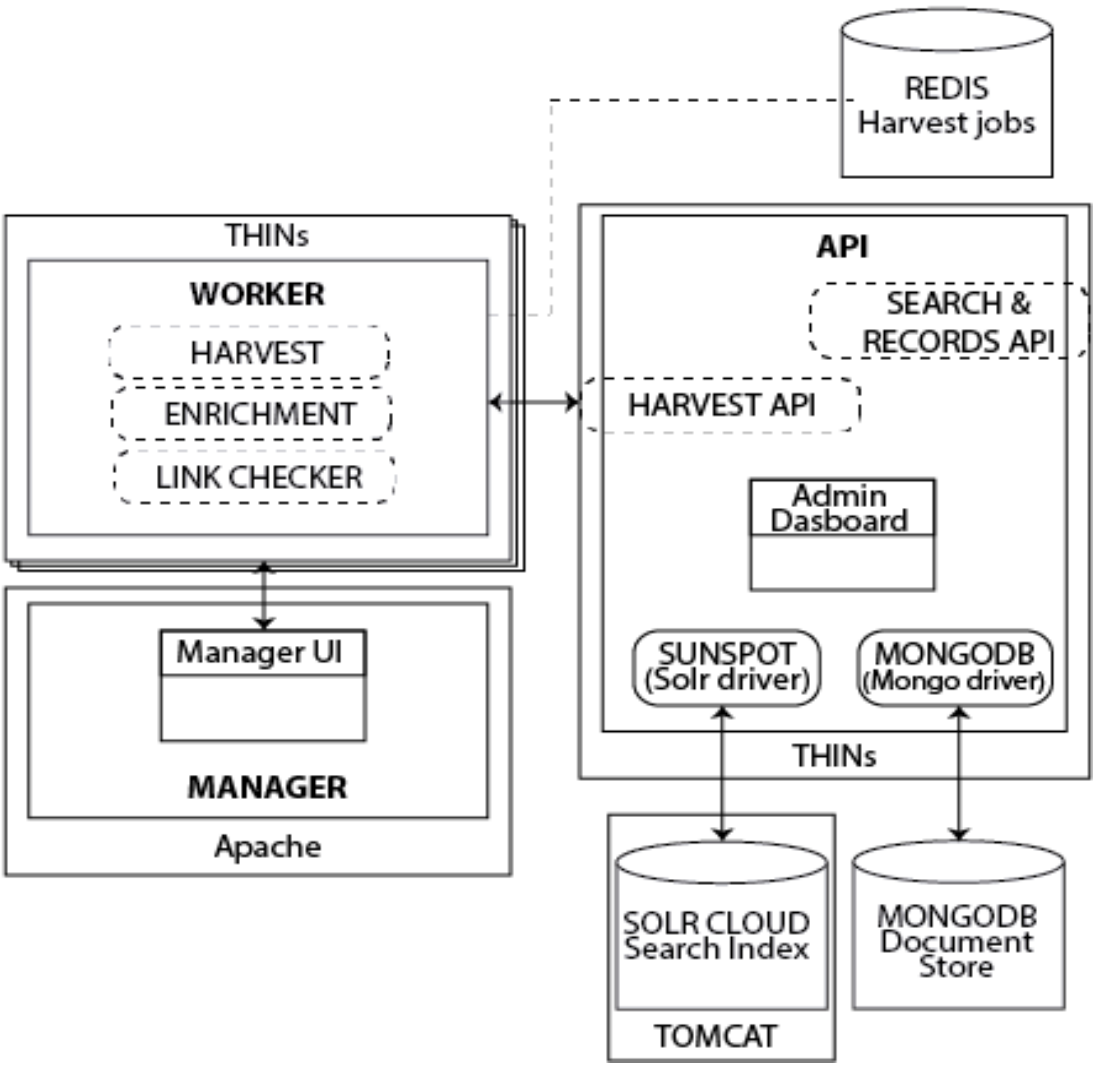
Fonte: Digital New Zealand (2018). Adaptada.

BIBLIOTECA NACIONAL DA NOVA ZELÂNDIA – DIGITALNZ

A Biblioteca Nacional da Nova Zelândia junto à Rede do povo Aotearoa Kaharoa desenvolveu, no início de 2006, o DigitalNZ, que utiliza o software Supplejack para agregação de dados (DIGITAL NEW ZEALAND, 2019). A figura 3, apresenta uma ilustração que demonstra como a agregação da DigitalNZ acontece.

A figura 3 mostra a ferramenta *Supplejack* como ferramenta central para agregação dos dados, dessa forma, a figura 4, apresenta a arquitetura da plataforma *Supplejack*.

Figura 4 – Arquitetura da plataforma Supplejack



Fonte: Supplejack (2020). Adaptada.

A arquitetura é composta por: *Manager* ou Gerenciador, que apresenta uma interface para o usuário controlar as atividades do software; *Worker* ou Trabalhador, que realiza as atividades de coleta, enriquecimento e verificação de links; API, um *wrapper* público para pesquisar o repositório de índice e metadados; *Common* ou Comum, são ajudantes compartilhados entre o *Worker* e o *Manager* (SUPPLEJACK, 2020).

Como apresentado na figura 4, o Supplejack depende da integração com um índice de pesquisa, o padrão é Solr, e um repositório de metadados, o padrão é MongoDB. O Apache SOLR trata-se de tecnologia voltada para pesquisa e indexação de documentos massivos, e MongoDB, um banco de dados do tipo NoSQL, também utilizado em projetos contemporâneos que envolvem novas arquiteturas para processamento de dados massivos baseados em informação semiestruturada ou mesmo desestruturada.

FUNDAÇÃO EUROPEANA – EUROPEANA

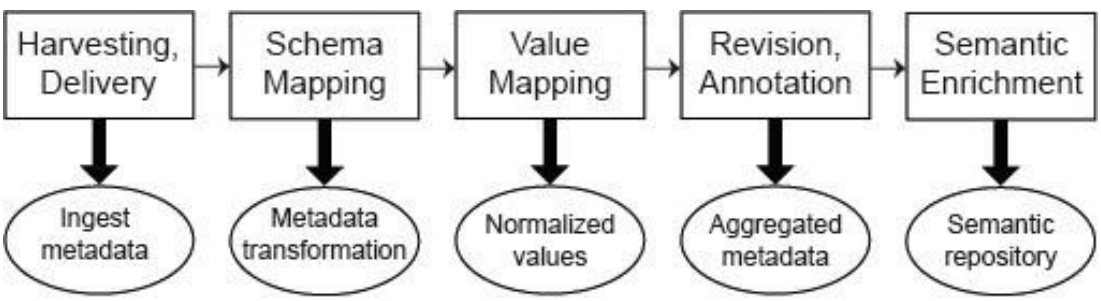
A Fundação Europeana desenvolveu a Europeana, que reuniu mais de 55 milhões de objetos digitais das coleções on-line de mais de 3.500 galerias, bibliotecas, museus, coleções audiovisuais e arquivos de toda a Europa (SCHOLZ, 2019). A figura 5 apresenta seu workflow de agregação.

A primeira etapa, "*Harvesting, Delivery*" ou "Colheita, Entrega", refere-se à coleta de metadados de provedores de conteúdo, por meio de protocolos de entrega, como o OAI-PMH, o HTTP e o FTP. A segunda etapa, "*Schema Mapping*", ou "Mapeamento de esquema", alinha os metadados coletados a um modelo de referência comum.

Nesta etapa, uma interface gráfica colabora com o mapeamento, por meio de uma linguagem de mapeamento compreensível por máquinas. A terceira etapa, "*Value Mapping*" ou "Mapeamento de valores", foca no alinhamento e na transformação dos termos constantes nos metadados coletados para arquivo de autoridade ou fonte externa, ou seja, permite a normalização de datas, localizações, países, idiomas, dentre outros.

A quarta etapa, "*Revision, Annotation*", ou "Revisão, Anotação", permite adicionar anotações para atribuir metadados não disponíveis no contexto original, e, por último, na quinta etapa, "*Semantic Enrichment*" ou "Enriquecimento semântico", foca na transformação dos dados em um modelo semântico, extração e identificação de recursos e implantação em RDF (SCHOLZ, 2019).

Figura 5 − Workflow de agregação de dados proposto pela Europeana



Fonte: Kollia *et al.*(2012, p. 70). Adaptada.

A documentação da Europeana não entra em detalhes tecnológicos específicos, não ficando claro que ferramentas e tecnologias são utilizadas em cada etapa, como as mesmas são parametrizadas e que esforços foram desenvolvidos para a integração dos serviços. Novamente, há dificuldade de se encontrar evidências que facilitem a replicação ou mesmo a adaptação de soluções em outros contextos.

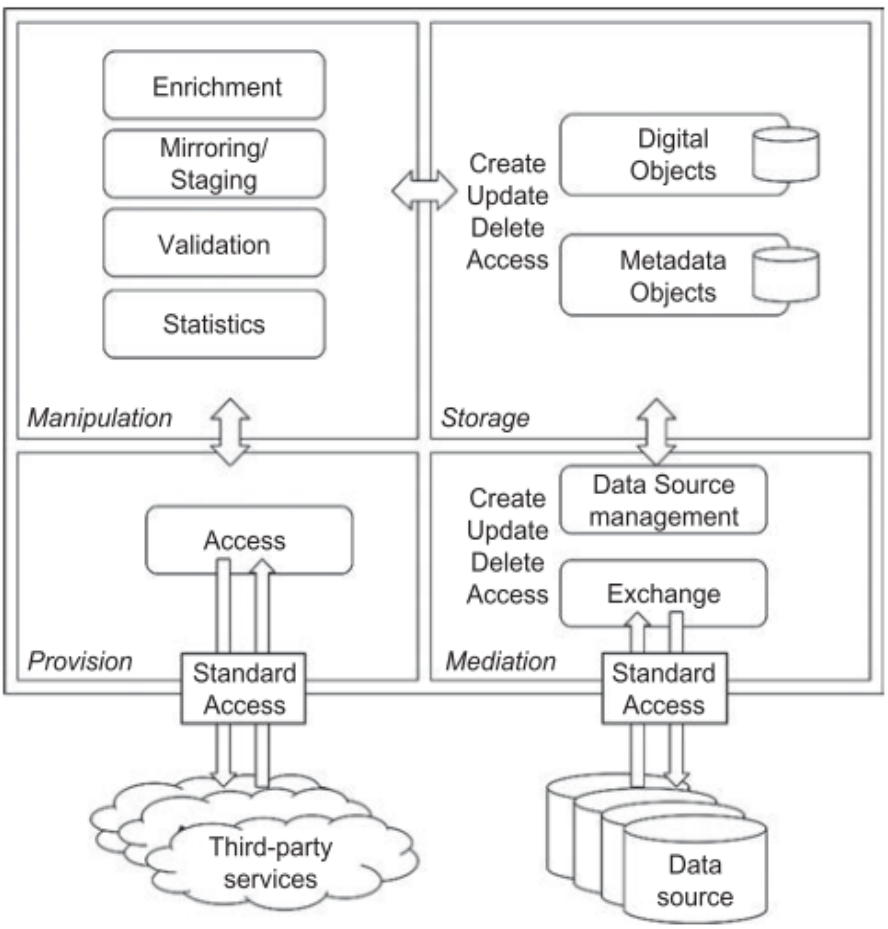
INSTITUTO DE CIÊNCIA E TECNOLOGIA DA INFORMAÇÃO - D-NET

O *Istituto di Scienza e Tecnologie dell'Informazione* desenvolveu o D-NET, uma estrutura orientada a serviços, de uso geral, na qual os designers podem construir infraestruturas agregadas autônomas, robustas, escaláveis e personalizadas, de maneira econômica.

Oferece serviços de gerenciamento de dados capazes de fornecer acesso a diferentes tipos de fontes de dados externas, armazenar e processar objetos de informações de qualquer modelo de dados, convertê-los em formatos comuns e expor objetos de informações a aplicativos de terceiros por meio de vários acessos padrão (MANGHI *et al.*, 2014).

Neste estudo, categorizamos D-NET e o Supplejack em um mesmo nicho, visto ambos serem softwares que têm por objetivo fornecer um serviço completo para agregação de dados. A figura 6 apresenta a arquitetura do software.

Figura 6 – Infraestrutura D-NET Software Toolkit



Fonte: Manghi *et al.* (2014, p. 327).

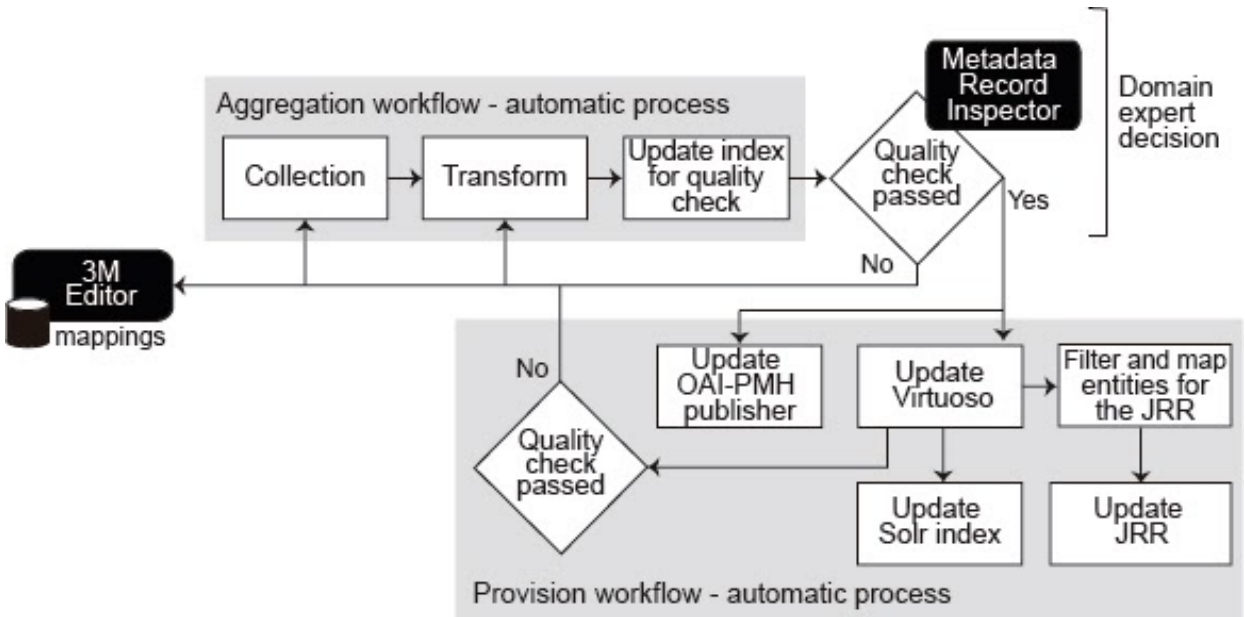
O D-NET subdivide sua arquitetura em 4 camadas principais: *Manipulation, Storage, Provision* e *Mediation*. Os serviços de manipulação, *Manipulation*, foram projetados para executar o enriquecimento, a validação, o espelhamento e a geração de estatísticas. Os serviços de armazenamento, *Storage*, como o próprio nome diz, propiciam o armazenamento de objetos, agrupando tecnologias conhecidas, de código-fonte aberto, como índices de texto completo, bancos de dados relacionais, repositórios de documentos etc. Os serviços de fornecimento de dados, *Provision*, fazem interface com aplicativos externos, como, por exemplo, portais para uso dos usuários finais ou serviços de terceiros. Além do acesso aleatório, o D-NET suporta as APIs: OAI-PMH *publisher service* e OAI-ORE *publisher service*. Os serviços de mediação, "*Mediation"*, visam a buscar dados de fontes externas e importá-los para a infraestrutura agregada, como, por exemplo, objetos em conformidade com um determinado recurso de modelo de dados (BARDI; MANGHI; ZOPPI, 2012).

Conforme supracitado, o Supplejack foi utilizado pela DigitalNZ. Nesse estudo, citamos o Parthenos Aggregator, que utiliza do D-NET.

INSTITUTO DE CIÊNCIA E TECNOLOGIA DA INFORMAÇÃO - PARTHENOS AGGREGATOR

As infraestruturas de humanidades digitais (DHIs) apoiam pesquisadores no campo das ciências humanas, oferecendo um ambiente digital, no qual podem encontrar e usar ferramentas e dados de pesquisa para conduzir suas atividades. Há um número crescente de DHIs e, para integrá-los, a Comissão Europeia lançou o projeto *Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies* (PARTHENOS) (FROSINI *et al.*, 2018). A figura 7 apresenta o *workflow* proposto.

Figura 7 – Workflow de agregação e provisão Parthenos Aggregator



Fonte: Frosini *et al.* (2018, p. 40). Adaptada.

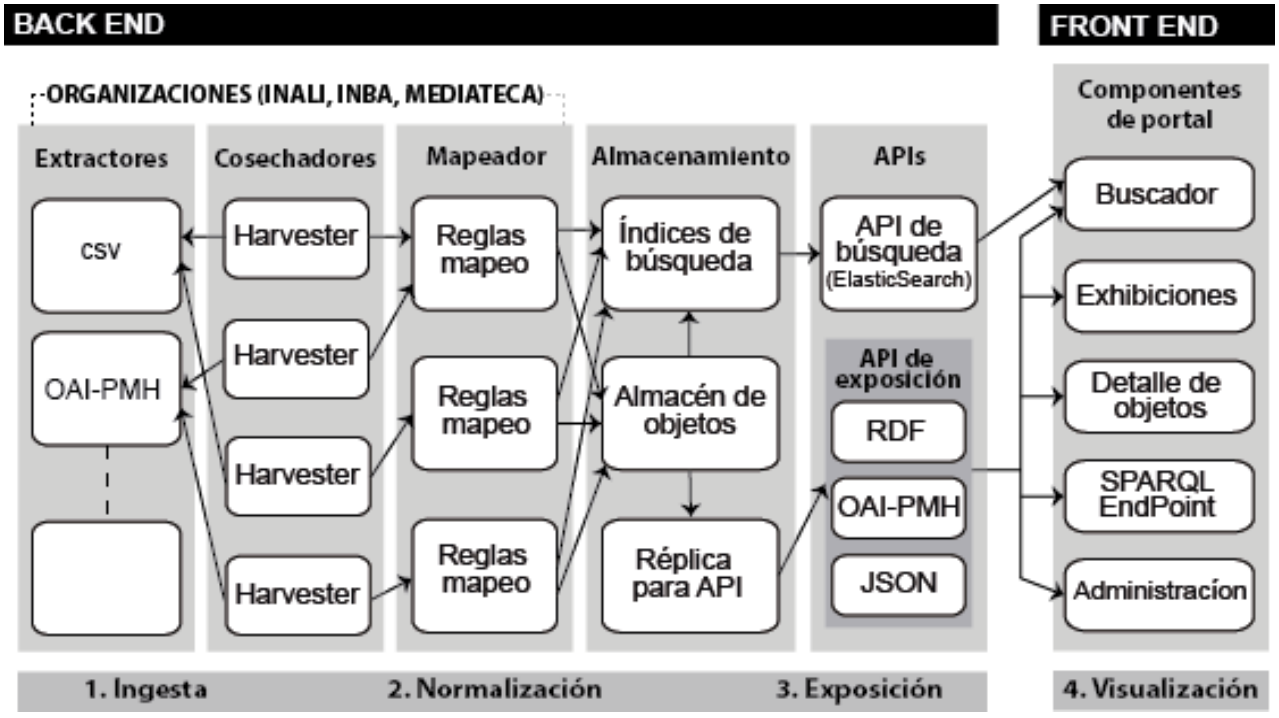
O *workflow* se divide em dois fluxos de trabalho: agregação e provisionamento. No fluxo de agregação, a etapa "*Collection*" ou "Coleta" visa a lidar com a coleta de metadados por meio de diferentes protocolos de acesso: OAI-PMH, FTP(S), SFTP, HTTP(S), RESTful. A etapa "*Transformation*" ou "Transformação" visa a mapear os metadados a uma ontologia única. A "*Metadata Cleaner*", ou "Limpeza de metadado", trata-se do serviço que harmoniza valores em registros de metadados com base em um conjunto de tesauros. Após esta etapa, inicia-se o processo de inspeção, na etapa "*Metadata Record Inspector*", ou "Inspeção de registro de metadados", na qual uma GUI da Web integrada ao D-NET, fornece dados aos curadores com uma visão geral das informações, possibilitando pesquisas e navegação entre os registros para verificar a correção da fase de transformação (por exemplo, metadados sem mapeamento, erros ou inconsistências semânticas) e a fase de limpeza.

Uma vez positivamente verificado, os registros podem ser exportados publicamente na "*OAI-PMH publisher service*". O serviço oferece Interfaces OAI-PMH para aplicativos de terceiros que desejam acessar metadados. "*Index service*" o serviço orienta a alimentação de Índices Solr e também é responsável por transformar os registros de metadados agregados em documentos Solr (FROSINI *et al.,* 2018).

SECRETARIA DE CULTURA DO MÉXICO – MEXICANA

A Secretaria de Cultura do México desenvolveu a Mexicana, um Repositório do Patrimônio Cultural do México, livre e aberto, que tem o objetivo principal de difundir e vincular os acervos do patrimônio cultural do México (MÉXICO, 2018). Seu *workflow* está apresentado na figura 8.

Figura 8 – Workflow de agregação Mexicana



Fonte: México (2018). Adaptada.

A Secretaria de Cultura do México desenvolveu documento explicativo sobre o projeto Mexicana, no qual consta o *workflow* apresentado, que é dividido em *Back End* e *Front End. No Back End,* a Etapa 1. *"Extractores*", ou "Extratores", trata-se dos componentes de código responsáveis pela cópia de dados locais ou remotos, realizando uma reestruturação mínima dos dados. Nesta etapa, considera-se a criação de extratores para diferentes formatos e a criação de interfaces para facilitar a extensão da funcionalidade.

A Etapa 2. "*Cosechadores*", ou "Coletores", trata-se de componente de código configurável que permite gerenciar os extratores e o mapeador dos dados (próxima etapa), de acordo com regras estabelecidas, tais como: execução sob demanda e por periodicidade. A Etapa 3., "*Mapeador*", permite configurar e executar regras de mapeamento definidas entre os dados gerados pelos extratores e o esquema de dados unificados do sistema.

A Etapa 4. "*Almacenamiento*", ou "Armazenamento", como o próprio nome diz, visa a armazenar os dados coletados. Nessa etapa, no *workflow*, há dois itens além do armazenamento: o "*Índices de búsqueda*", ou "Índices de pesquisa", no qual o componente explora os serviços do ElasticSearch para gerar índices de pesquisa de metadados de objetos incorporados ao armazenamento do sistema e a "*Réplica para API*", que replica os objetos digitais para fornecer seu acesso rápido através do APIs de exposição. A Etapa 5. "API", reforça o uso servidor ElasticSearch para indexação e recuperação de metadados, assim como a "API de Búsqueda" ou "API de Exposição" trata sobre formatos para exibição de objetos digitais e seus metadados.

No *Front end,* etapa voltada para visualização dos dados, trata-se do "Buscador" ou "Pesquisa", que trata sobre pesquisa desagregada de objetos digitais inseridos no sistema através da exploração da API; "*Exhibiciones*", ou "Exposições", que permite a configuração de coleções de objetos digitais agrupados por tema ou evento específico; "*Detalle de objetos*", ou "Detalhe do objeto", que permite visualizar o arquivo detalhes dos objetos digitais inseridos no sistema e no metadados associados; o "*EndPoint SPARQL*", que permite a execução de consultas SPARQL para o modelo de dados do sistema unificado através da estrutura semântica definida pelo Modelo de Dados e, por fim, "*Administración*", ou "Administração", que permite a administração de usuário e fluxo de trabalho para a configuração e execução das coletas para fontes de dados.

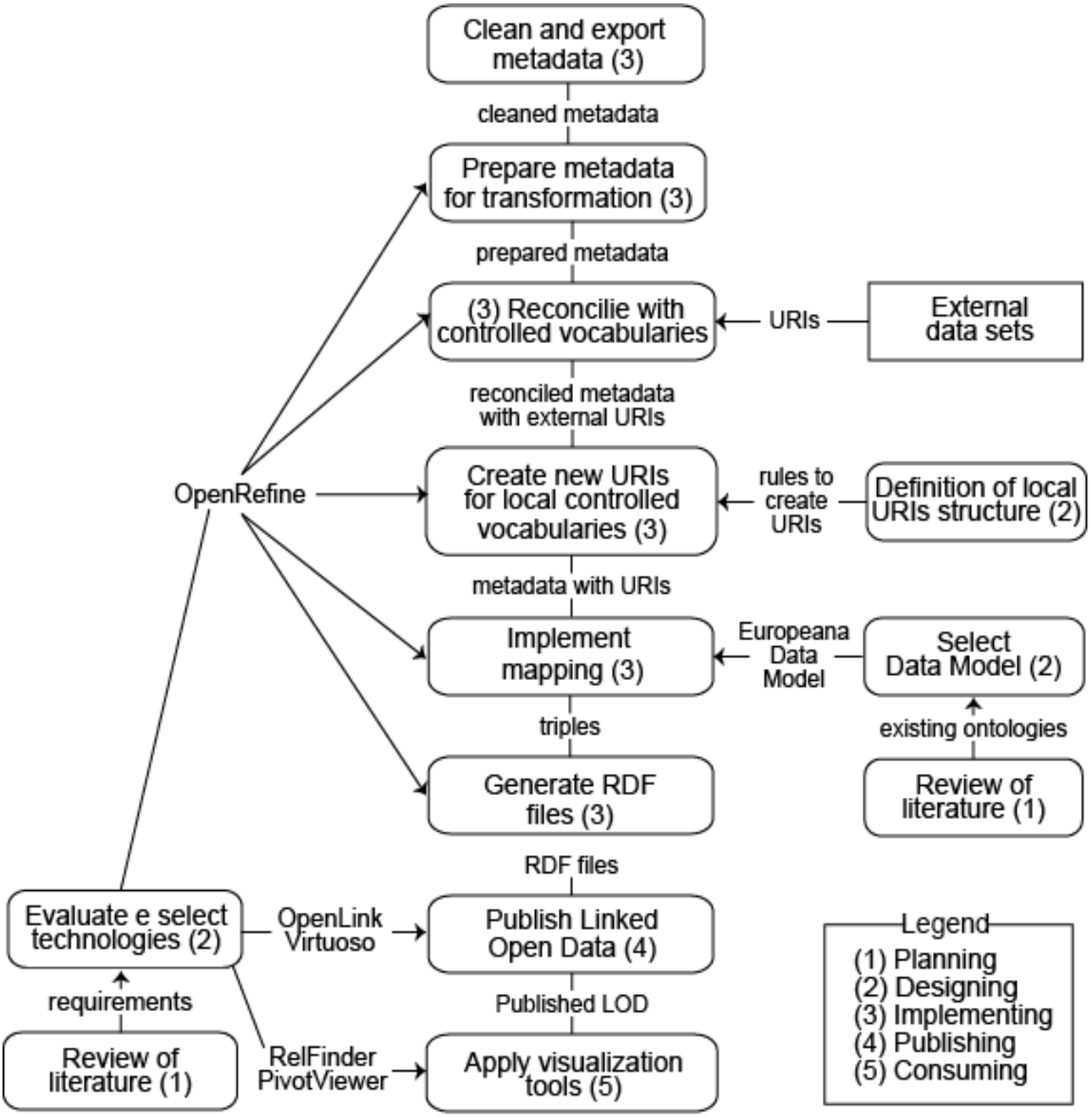
UNIVERSIDADE DE NEVADA - UNLV'S LINKED DATA PROJECT

A Universidade de Nevada, por meio da equipe do departamento de Coleções Digitais das Bibliotecas da Universidade, reuniu esforços para encontrar maneiras de tornar mais eficiente a descoberta e o uso das informações, iniciando, assim, estudos para adoção do *Linked Open Data* (LOD), culminando no desenvolvimento do *UNLV's Linked Data Project* (SOUTHWICK, 2015). A figura 9 apresenta o *workflow*.

Southwick (2015) afirma que, para o desenvolvimento do projeto, optou-se pela adoção de tecnologias com código aberto, sem qualquer adaptação ou desenvolvimento, ou seja, "no estado em que se encontram".

O projeto foi dividido em cinco etapas*: Planning, Designing, Implementing, "Publishing LOD" e "Consuming LOD"*, ou seja, Planejamento, Concepção, Implementação, "Publicando LOD" e "Consumindo LOD". O Planejamento é composto por duas revisões de literatura e retrata o período de estudo necessário ao desenvolvimento do protótipo.

Figura 9 – Workflow de agregação de dados proposto pela Universidade de Nevada



Fonte: Southwick (2015, p. 13). Adaptada.

A segunda etapa, Concepção, se desdobra em três atividades: "*Evaluate e select technologies*" ou "Avaliar e selecionar as tecnologias", na qual a equipe do projeto avaliou várias tecnologias e selecionou seis para aplicação no protótipo. A autora destaca que, embora as tecnologias selecionadas tenham funcionado bem, não significa que são as únicas ou as melhores; "*Select data model"* ou "Selecionar modelo de dados", que diz respeito à seleção da ontologia utilizada, a partir da investigação de modelos de dados utilizados por outras instituições; e "*Definition of local URIs strutucture*" ou "Definir a estrutura das URIs", fase na qual optou-se por criar URIs apenas para "coisas" que ainda não receberam URIs de outros provedores de dados. Se posteriormente forem descobertas URIs diferentes atribuídas à mesma "coisa" à qual já atribuímos um URI, elas são adicionadas ao conjunto de triplas que indicam a equivalência entre URIs.

A terceira fase, a Implementação, é composta por seis etapas: "*Clean and export metadata*" ou "Limpar e exportar metadados": o CONTENTdm é utilizado como sistema de gerenciamento de conteúdo e este exporta metadados em formato de planilha delimitada por tabulação, que pode ser importado para o OpenRefine. A limpeza consiste em cumprir rigorosamente os termos usados nas coleções que adotam um determinado vocabulário controlado e criar vocabulários controlados locais consistentes; "*Prepare metadata for transformation*", ou "Preparar metadados para transformação", fase de preparação dos metadados para gerar LOD. Para tal, utilizaram-se funções OpenRefine, como: remover espaços em branco; separar tipos diferentes de dados e separar valores agrupados em um único campo; "*Reconcilie with controlled vocabularies*", ou "Reconciliar com vocabulários controlados", trata-se da reconciliação feita usando a extensão "OpenRefine RDF", do OpenRefine e "*Generate RDF files*", ou "Gerar arquivos RDF", como o próprio nome diz, trata-se da geração de arquivos RDF que serão utilizados nas próximas etapas. As etapas "*Create new URIs for local controlled vocabularies*" e "*Implement mapping*" são a efetivação das etapas "*Definition of local URIs structure*" e "*Select data model*", da fase da Concepção.

A quarta fase, a Publicação, é composta por uma única atividade, "*Publish Linked Open Data*", ou "Publicar no Linked Open Data – LOD", na qual, o arquivo RDF é publicado à comunidade. A última fase, Consumo, também é composta por uma única atividade, "*Apply visualization tools*", ou "Aplicar ferramentas de visualização", na qual foram realizados três experimentos com ferramentas de visualização para arquivos RDF: com Pivot Viewer que foi útil para visualizar imagens de maneira muito dinâmica, pois é baseado em consultas SPARQL; com RelFinder, que visa a encontrar relacionamentos entre as "coisas" e também com RelFinder, mas com o conceito de relacionamento expandido, considerando relacionamentos que duas "coisas" tinham uma ou mais "coisas" em comum.

DISCUSSÃO DOS RESULTADOS

As interfaces de busca integradas estão disponíveis na rede e o quadro 2 apresenta o link para cada uma delas.

Quadro 2 – Links das interfaces de busca integradas

Projeto	Sites
AAC	https://americanart.si.edu/search
Trove	https://trove.nla.gov.au/
DigitalNZ	https://digitalnz.org/
Europeana	https://www.europeana.eu/pt/collections
D-NET Software	-
Parthenos Aggregator	http://www.parthenos-project.eu/portal
Repositório Mexicana	https://mexicana.cultura.gob.mx/
UNLV's Linked Data Project	https://www.library.unlv.edu/linked-data

Fonte: Elaborado pelos autores (2020).

Considerando a análise de todos os *workflows* apresentados acima, foram encontradas oito fases para agregação, sendo elas: extração, estruturação, transformação, reconciliação, armazenamento, exposição, publicação e novas aplicações. De forma sintética, estas etapas significam:

1. Extrair: extração dos dados em sua forma bruta, que podem estar, por exemplo, em pdf, em planilhas eletrônicas, documentos de texto, XML (*eXtensible Markup Language*), em bancos de dados relacionais, dentre outras opções.
2. Estruturar: selecionar vocabulários controlados pré-existentes e ontologias para aplicação nos dados.
3. Transformar: realizar a normalização, limpeza e correção sintática dos dados.
4. Reconciliar: enriquecer os metadados por meio de outros dados existentes na web.
5. Armazenar: se trata da escolha de onde os dados coletados serão armazenados.
6. Publicar: se trata da interface única de busca integrada.
7. Expor: disponibilizar os dados agregados por meio de API, que exponham os dados em formato RDF, OAI-PMH ou JSON.
8. Possibilitar novas aplicações: a partir dos arquivos disponibilizados na etapa 'Expor', considerar que novas aplicações podem ser criadas.

O quadro 3 mostra um resumo individual, considerando a presença (X) ou não (-) de cada etapa, para visualização geral.

Quadro 3 − Etapas dos Workflows de Agregação, panorama individual

Projeto/Etapas	Extrair	Estruturar	Transformar	Reconciliar	Armazenar	Publicar	Expor	Novas aplicações
AAC	X	X	-	X	X	X	X	X
Trove [1]	X	-	-	-	X	X	-	-
DigitalNZ	X	-	-	-	X	X	-	-
Europeana [2]	X	X	-	X	X[2]	X[2]	-	-
D-NET Software	X	X	X	X	X	X	X	X
Parthenos Aggregator	X	X	X	X	X	X	X	X
Repositório Mexicana	X	X	-	X	X	X	X	-
UNLV's Linked Data Project	X	X	X	X	X	X	X	X
[1] Dados observados somente a partir da visualização dos workflows [2] Itens não explícitos no workflow, mas identificados na documentação								

Fonte: Elaborado pelos autores (2020).

O quadro 4 apresenta a nomenclatura original das etapas na literatura revisada, classificando-as dentro das etapas elencadas neste estudo. Este quadro também nos ajuda a visualizar quais são os nomes mais usados para nomear cada etapa, para colaborar com pesquisas futuras.

A documentação na qual os *workflows* estão inseridos apresentam alguns dados que não constam do fluxograma. Além disso, percebe-se pouca preocupação com a qualidade dos dados inseridos, ou seja, os dados coletados na etapa de extração, havendo pouca menção a etapas tradicionais de projetos de análise de dados, envolvendo limpeza, tratamento e normalização de dados.

Além das etapas, as publicações apresentam algumas ferramentas de software utilizados para execução do *workflow*. De forma geral, os *workflows* são genéricos demais e não apresentam o fluxo real de processos necessários, contrariando assim, um dos princípios básicos de um *workflow*, que é a possibilidade de ser replicado. Além disso, percebe-se a necessidade de um conhecimento técnico avançado e extremamente especializado para compreensão de todas as etapas.

CONCLUSÕES

A análise dos diferentes *workflows* de agregação de dados permitirá aos pesquisadores compreender quais etapas estão sendo executadas, quais estão sendo postas em segundo plano e quais precisam ser incluídas. Esse conhecimento estruturado pode auxiliar na compreensão de etapas que devem ser resolvidas do ponto de vista da criação de um serviço de agregação de dados culturais. Além disso, é importante compreender que não há consenso nem na quantidade de etapas, no seu nome e nem nas tecnologias utilizadas, demonstrando o quanto esse tema parece ainda em estágio inicial de pesquisa ou mesmo revelando que as soluções são altamente customizadas, exigindo soluções locais para problemas específicos.

É importante destacar que a grande maioria menciona soluções para processamento de dados massivos e construídas para lidar com projetos de *big data*. São mencionados o Apache Lucene, Apache Solr, ElasticSearch, Hadoop, MapReduce e MongoDB. Não fica clara a forma como essas tecnologias são utilizadas, a maneira como são integradas e a documentação se mostra bastante deficitária de detalhes e discussões alongadas sobre o tema. No entanto, é importante perceber que já há na discussão sobre a agregação de dados culturais a presença dessas tecnologias de forma determinante. É importante reconhecer que esse é um tema ainda novo para a Ciência da Informação e que esforços de pesquisa e desenvolvimento devem ser feitos para que se compreendam as possíveis aplicações dessas tecnologias, dado que as mesmas não apenas são novas técnicas, mas representam novas formas de se pensar nos dados e em um ecossistema completo de serviços analíticos.

Também é possível notar a baixa densidade dos trabalhos apresentados, sendo as discussões feitas de forma bastante genérica. O trabalho mais detalhado identificado diz respeito à iniciativa menos automatizada, relacionada ao trabalho da Universidade de Nevada (SOUTHWICK, 2015), que fez intensivo da ferramenta OpenRefine como estratégia de coleta, tratamento e organização dos dados. Apesar da importância da pesquisa, a mesma demonstra que todo o fluxo de trabalho deveria ser feito novamente para cada novo registro publicado, inviabilizando sua adoção para solução para serviços que exigem atualização automática dos índices de busca e recuperação da informação.

Fica evidente, a partir dos resultados desta pesquisa, o quanto ainda é necessário se compreender como esses fluxos de agregação devem funcionar e como podem ser utilizados para a criação de serviços informacionais de agregação de dados. Vale ressaltar que serviços dessa ordem representam grandes contribuições da área da Ciência da Informação para a sociedade brasileira, assim como tem sido com serviços como a Biblioteca Digital de Teses e Dissertações (BDTD) criada pelo IBICT e a própria BRAPCI, no caso específico da comunidade da Ciência da Informação.

Como trabalho futuro, pretende-se realizar pesquisas direcionadas a cada etapa do *workflow*, buscando ampliar a compreensão de como as etapas são realizadas, seus detalhes operacionais, técnicos e informacionais.

Quadro 4 – Nomenclatura original das etapas na literatura revisada

Projeto/ Etapas	**Extrair**	**Estruturar**	**Transformar**	**Reconciliar**	**Armazenar**	**Publicar**	**Expor**	**Novas aplicações**
AAC	*Prepare and export*	*Define AAC Model*	-	*Reconcilie Entities*	*RDF Triple Store - SPARQL*	*Browse Demo APP e Toy Box Applets*	*SPARQL/ Transform Library API*	*Future Applications*
Trove [1]	*NLA Harvest*	-	-	-	*MySQL Seachable Unit Database*	*Trove User Interface*	-	-
DigitalNZ	*Manager e Common*	-	-	-	*API*	*API*	-	-
Europeana	*Harvesting Delivery*	*Schema Mapping*	-	*Value Mapping*	*X* [2]	*X* [2]	-	-
D-NET Software	*Mediation*	*Manipulation*	*Manipulation*	*Manipulation*	*Storage*	*Provision*	*Provision*	*Provision*
Parthenos Aggregator	*Collection*	*Transformation*	*Metadata Record Inspector*	*Metadata Cleaner*	*Index Service*	*X* [2]	*OAI-PMH publisher service*	*OAI-PMH publisher service*
Repositório Mexicana	*Extractores e Cosechadores*	*Mapeador*	-	-	*Almacienamento*	*Buscador e Exhibiciones*	*API de Búsqueda*	-
UNLV's Linked Data Project	*Clean and export*	*Implement mapping*	*Prepare matadata for transformation*	*Reconcilie with controlled vocabularies e Create new URIs for local controlled vocabularies*	*Publish LOD*	*Publish LOD*	*Apply visualization tools*	*Publish LOD*
[1] Dados observados somente a partir da visualização do workflow								
[2] Item não explicito no workflow								

Fonte: Elaborado pelos autores (2020).

REFERÊNCIAS

BARDI, A.; MANGHI, P.; ZOPPI, F. Aggregative Data Infrastructures for the Cultural Heritage. *In*: DODERO, J. M.; PALOMO-DUARTE, M.; KARAMPIPERIS, P. (org.). *Metadata and Semantics Research*. Communications in Computer and Information Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. v. 343, p. 239–251. DOI 10.1007/978-3-642-35233-1_24. Disponível em: http://link.springer.com/10.1007/978-3-642-35233-1_24. Acesso em: 11 mar. 2021.

BRIGHAM, T. J.; FARRELL, A. M.; OSTERHAUS TRZASKO, L. C.; ATTWOOD, C. A.; WENTZ, M. W.; ARP, K. A. Web-Scale Discovery Service: Is It Right for Your Library? Mayo Clinic Libraries Experience. *Journal of Hospital Librarianship*, v. 16, n. 1, p. 25–39, 2 jan. 2016. DOI 10.1080/15323269.2016.1118280. Disponível em: http://www.tandfonline.com/doi/full/10.1080/15323269.2016.1118280. Acesso em: 11 mar. 2021.

DIGITAL NEW ZEALAND. *Our History*. 2019. Disponível em: https://digitalnz.org/about/our-history. Acesso em: 18 abr. 2020.

DIGITAL NEW ZEALAND. This is Digital New Zealand. 20 dez. 2018. *YouTube*. Disponível em: https://www.youtube.com/watch?v=UWbIDwsaA4o. Acesso em: 18 abr. 2020.

EUROPEANA. *Brief History*. 2020. Disponível em: https://pro.europeana.eu/about-us/mission#brief-history. Acesso em: 27 abr. 2020.

FINK, E. E. Overview and Recommendations for Good Practices. *American Art Collaborative. Linked Open Data Initiative*, 2018. Disponível em: http://americanartcollaborative.org/wp-content/uploads/2018/03/AAC_LOD_Overview_Recommendations.pdf. Acesso em: 15 abr. 2020.

FROSINI, L.; BARDI, A.; MANGHI, P.; PAGANO, P. An Aggregation Framework for Digital Humanities Infrastructures: The PARTHENOS Experience. *SCIentific RESearch and Information Technology*, v. 8, n. 1, 11 jul. 2018. DOI 10.2423/i22394303v8n1p33. Disponível em: https://doi.org/10.2423/i22394303v8n1p33. Acesso em: 11 mar. 2021.

KOLLIA, I.; TZOUVARAS, V.; DROSOPOULOS, N.; STAMOU, G. A systemic approach for effective semantic access to cultural content. *Semantic Web*, v. 3, n. 1, p. 65–83, 2012. DOI 10.3233/SW-2012-0051. Disponível em: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-2012-0051. Acesso em: 11 mar. 2021.

MANGHI, P.; ARTINI, M.; ATZORI, C.; BARDI, A.; MANNOCCI, A.; LA BRUZZO, S.; CANDELA, L.; CASTELLI, D.; PAGANO, P. The D-NET software toolkit: A framework for the realization, maintenance, and operation of aggregative infrastructures. *Program*, v. 48, n. 4, p. 322–354, 27 ago. 2014. DOI 10.1108/PROG-08-2013-0045. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/PROG-08-2013-0045/full/html. Acesso em: 11 mar. 2021.

MÉXICO. SECRETARÍA DE CULTURA. *Mexicana Repositorio del Patrimonio Cultural de México*. Ciudad de México: Secretaría de Cultura, 2018. Disponível em: https://mexicana.cultura.gob.mx/work/models/repositorio/Resource/126/2/images/documentacion.pdf. Acesso em: 18 abr. 2020.

NATIONAL LIBRARY OF AUSTRALIA. *Trove Help Center*. Trove System Architecture Diagram. 2010. Disponível em: https://www.nla.gov.au/trove/marketing/Trove%20architecture%20diagram.pdf. Acesso em: 18 abr. 2020.

NAVARRETE, T. *Europeana as online cultural information service: study report.* [*S. l.*]: Europeana, 2016. Disponível em: https://pro.europeana.eu/files/Europeana_Professional/Publications/europeana-benchmark-report-sep-2016.pdf. Acesso em: 27 abr. 2020.

PAVÃO, C. M. G.; CAREGNATO, S. E. Serviços de descoberta em rede: a experiência do modelo Google para os usuários de bibliotecas universitárias. *Em Questão,* v. 21, n. 3, p. 130, 2015. Disponível em: https://seer.ufrgs.br/EmQuestao/article/view/58410/36046. Acesso em: 27 abr. 2020.

PEREIRA, L. A. M.; CASANOVA, M. A. *Sistemas de gerência de workflows: características, distribuição e exceções.* Rio de Janeiro: PUC-Rio, 2003. Disponível em: ftp://ftp.inf.puc-rio.br/pub/docs/techreports/03_11_pereira.pdf. Acesso em: 27 ago. 2020.

SCHOLZ, H. *A guide to the metadata and content requirements for data partners publishing material in Europeana Collections.* [*S. l.*]: Europeana Foundation, 2019(Europeana Publishing Guide, v1.8). Disponível em: https://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20Publishing%20Guide%20v1.8.pdf. Acesso em: 18 abr. 2020.

SOUTHWICK, S. B. A Guide for Transforming Digital Collections Metadata into Linked Data Using Open Source Technologies. *Journal of Library Metadata*, v. 15, n. 1, p. 1–35, 2 jan. 2015. DOI 10.1080/19386389.2015.1007009. Disponível em: http://www.tandfonline.com/doi/abs/10.1080/19386389.2015.1007009. Acesso em: 11 mar. 2021.

SUPPLEJACK. *Architecture. Documentation* (Version 0.1). 2020. Disponível em: http://digitalnz.github.io/supplejack/architecture.html. Acesso em: 18 abr. 2020.

TROVE HELP CENTRE. *About Trove*. 2020. Disponível em: https://help.nla.gov.au/trove/using-trove/getting-to-know-us. Acesso em: 18 abr. 2020.

AGRADECIMENTOS

Agradecimento ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPQ que financiou a pesquisa por meio da bolsa de doutorado.

Publicando dados de pesquisa: contextualizando as principais etapas e elementos envolvidos no processo

Guilherme Ataíde Dias
Pós-Doutorado pela Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp) - Brasil. Doutor em Ciências da Comunicação /Ciência da Informação pela Universidade de São Paulo (USP) - Brasil. Professor da Universidade Federal da Paraíba (UFPB) - João Pessoa, PB - Brasil
http://lattes.cnpq.br/9553707435669429
https://orcid.org/0000-0001-6576-0017
E-mail: guilhermeataide@ccsa.ufpb.br

Sandra de Albuquerque Siebra
Doutora em Ciências da Computação pela Universidade Federal de Pernambuco (UFPE) – PE - Brasil. Professora da Universidade Federal de Pernambuco (UFPE) - Recife, PE - Brasil
http://lattes.cnpq.br/4923627544089379
https://orcid.org/0000-0002-0078-6918
E-mail: sandra.siebra@gmail.com

Rosilene Paiva Marinho de Sousa
Doutora em Ciência da Informação pela Universidade Federal da Paraíba (UFPB) – PB - Brasil. Professora da Universidade Federal do Oeste da Bahia (UFOB) - Barreiras, BA – Brasil.
http://lattes.cnpq.br/4465533418771961
https://orcid.org/0000-0002-4699-8692
E-mail: adv.rpmarinho@gmail.com

Marckson Roberto Ferreira de Sousa
Doutor em Engenharia Elétrica pela Universidade Federal da Paraíba (UFPB) – PB - Brasil. Professor da Universidade Federal da Paraíba (UFPB) - João Pessoa, PB – Brasil.
http://lattes.cnpq.br/0221265788966967
https://orcid.org/0000-0003-2001-1631
E-mail: marckson.dci.ufpb@gmail.com

Submetido em: 24/11/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021 .

RESUMO

O uso e reúso de dados de pesquisa são ações importantes e necessárias no contexto atual do fazer científico. Por isso, os conjuntos de dados produzidos durante as iniciativas de pesquisa precisam ser publicados para serem mais facilmente acessados por toda a comunidade científica. Porém, publicar dados não significa a mera disponibilização destes em um repositório, mas compreende uma série de etapas que devem ser consideradas para que eles possam, efetivamente, serem utilizados. Assim, este trabalho tem como objetivo principal discutir o processo de publicação de conjuntos de dados de pesquisa no contexto da ciência. Esta é uma investigação descritiva e qualitativa, que faz uso da pesquisa bibliográfica. Como resultado são apresentados elementos envolvidos no processo de publicação de dados de pesquisa e discussões abrangendo as fases de publicação de conjuntos de dados, relativas ao depósito, descrição, atribuição de identificador e revisão, etapas essas que envolvem o núcleo do processo de publicação.

Palavras-chave: Publicação de dados. Compartilhamento de dados. Artigo de dados. Periódico de dados. Dados científicos.

Publishing research data: contextualizing the main steps and elements involved in the process

ABSTRACT

The use and reuse of research data are important and necessary actions in the current context of scientific practice. For this reason, the datasets produced during the research initiatives need to be published to be more easily accessed by the entire scientific community. However, publishing data does not mean merely making it available in a repository, but comprises a series of steps that must be considered so that they can be effectively used. Thus, this work has as main objective to discuss the process of publishing research datasets in the context of science. This is a descriptive and qualitative investigation, which makes use of bibliographic research. As results, elements involved in the process of publishing research data and discussions are presented, covering the stages of publication of datasets, relating to the deposit, description, identifier assignment and review, steps that involve the core of the publication process.

Keywords*: Data publication. Data sharing. Data paper. Data journal. Scientific data.*

Publicación de datos de investigación: contextualización de los principales pasos y elementos implicados en el proceso

RESUMEN

El uso y reúso de datos de investigación son acciones importantes y necesarias en el contexto actual del hacer científico. Por esta razón, los conjuntos de datos producidos durante las iniciativas de investigación necesitan ser publicados, de tal manera que, toda la comunidad científica pueda acceder fácilmente a ellos. Sin embargo, publicar datos no significa simplemente colocarlos a disposición en un repositorio, sino comprende una serie de fases que deben ser consideradas para que, los datos, puedan utilizarse de manera efectiva. Es así que, este estudio tiene como objetivo principal discutir el proceso de publicación de conjuntos de datos de investigación en el contexto de la ciencia. Se trata de una investigación de tipo descriptiva con enfoque cualitativo, la cual hace uso de investigación bibliográfica. Como resultado, se presentan elementos involucrados en el proceso de publicación de los datos de investigación, y discusiones que abarcan las fases de publicación de los conjuntos de datos, relativas al depósito, la descripción, la asignación del identificador y la revisión, fases que constituyen el núcleo del proceso de publicación.

Palabras clave*: Publicación de datos. Uso compartido de datos. Artículo de datos. Revista de datos. Datos científicos.*

INTRODUÇÃO

Compartilhar dados de pesquisa é uma ação entendida como de fundamental importância pela comunidade científica (BORGMAN, 2015; DRAZEN et al, 2016). Os benefícios associados ao compartilhamento de dados, tais como a redução de custos no processo de investigação; a reprodutibilidade dos experimentos; a possibilidade de comprovação de resultados obtidos; o aumento das colaborações entre pesquisadores e o retorno para a sociedade dos investimentos públicos realizados através de órgãos de fomentos, contribuindo para um melhor impacto socioeconômico das pesquisas públicas, dentre outros, são bastante evidentes. Porém, para os dados de pesquisa terem seu uso potencializado por meio do compartilhamento, carecem de uma publicação eficaz, o que significa não simplesmente disponibilizá-los no servidor de um programa de pós-graduação ou de uma instituição de pesquisa, na página pessoal do próprio pesquisador, ou mesmo em um repositório institucional de uma instituição de pesquisa ou universidade. O processo de publicação de dados de pesquisa pode influenciar no efetivo compartilhamento destes recursos e na ampliação ou não do seu uso e reúso. Neste trabalho adota-se a definição de Publicação de dados (com P maiúsculo) de Callaghan *et al.* (2012), que aponta que a Publicação é um processo formal que deve: fornecer mecanismo para assegurar o crédito ao trabalho do pesquisador (de forma que os dados possam ser formalmente citados); possibilitar que se agregue valor a um conjunto de dados (pois estes precisam estar documentados) e garantir a persistência destes.

Para que a Publicação de dados de pesquisa seja bem sucedida é necessário o envolvimento e conscientização e, às vezes, a capacitação/treinamento da comunidade de pesquisa no processo. E para a motivação desta comunidade, é importante divulgar para os pesquisadores as vantagens que eles podem obter ao adotar a referida prática. Costello (2009) lista uma série de benefícios associados à publicação de dados que podem ser auferidos ao cientista individual responsável pela criação de conjuntos de dados. Estes benefícios, dentre outros, incluem: "Maior taxa de citação"; "Reconhecimento entre os pares ampliado"; "Convites para encontros"; "Convites para publicar"; "Convites para prover consultoria"[1] (COSTELLO, 2009, p. 420), além de se criar oportunidades para parcerias e trabalhos colaborativos.

Modelos de ciclo de vida de dados tem se mostrado úteis para orientar, definir e ilustrar visualmente os processos/etapas pela quais os dados devem passar até a sua Publicação. Estes ciclos englobam atividades/ações a serem realizadas, de acordo com as necessidades específicas de cada pesquisador e os tipos de dados sendo trabalhados, além de papéis e responsabilidades, marcos e outros componentes importantes para a gestão, preservação e disponibilização dos dados (CARLSON, 2014; KOWALCZYK, 2018). Existem diversos modelos de ciclo de vida dos dados, tais como o DCC Curation Lifecycle Model (DCC, 2020), o DataONE Data Life Cycle (DATAONE, 2020), o DDI Combined Life Cycle Model e o Research Data Lifecycle – UK Data Service (UK DATA SERVICE, 2020). Pode ser observado que esses modelos apresentam diferentes níveis de granularidade e detalhes, mas todos são organizados em processos/etapas que possuem semelhanças, sendo as principais: planejamento; criação, coleta ou captura; armazenamento; gestão a longo prazo/preservação; processamento ou análise; Publicação ou compartilhamento; uso e reúso. (CORTI *et al.*, 2014; KOWALCZYK, 2018). Entretanto, estes modelos, se observados sob a ótica dos pesquisadores, podem ser abstratos e apresentar uma certa complexidade para orientar o processo de Publicação dos dados. Assim, um modelo mais recente encontrado na literatura, focado na Publicação de dados, é o apresentado por Kratz e Strasser (2014).

[1] Texto original em inglês:
"Greater citation rate"
"Wider recognition among peers"
"Invitations to meetings"
"Invitations to collaborate"
"Invitations to provide consultancy"

Um modelo que simplifica o processo de Publicação dos conjuntos de dados e, apesar de não englobar todas as etapas relevantes dos modelos de ciclo de vida de dados existentes (tais como os anteriormente citados), tem potencial para deixar mais claro e simples o processo de Publicação de dados para os pesquisadores em geral.

Nesse contexto, o objetivo principal deste trabalho é discutir o processo de Publicação de conjuntos de dados de pesquisa, tomando como referência o modelo apresentado por Kratz e Strasser (2014).

As pesquisas de Dias, Anjos e Araújo (2019) e de RDP BRASIL (2018) revelaram que muitos dos pesquisadores brasileiros de todas as áreas do conhecimento possuem incertezas sobre questões relacionadas aos dados de pesquisa. Assim, espera-se, com esse trabalho, contribuir para o aumento da compreensão da temática por cientistas de todas as áreas do conhecimento, assim como com o despertar para outras questões associadas aos dados de pesquisa.

METODOLOGIA

Do ponto de vista de seu objetivo, esta pesquisa é descritiva. Para Richardson (2017, p. 8), a pesquisa descritiva "[...] procura descrever sistematicamente uma situação, problema, fenômeno ou programa para revelar da estrutura o comportamento de um fenômeno".

Os recursos bibliográficos usados na investigação foram obtidos por meio de consultas diretas e indiretas (consulta às referências dos materiais obtidos) realizadas no Portal de Periódicos da CAPES. Este portal foi selecionado em virtude da possibilidade de acessar de maneira centralizada diversas bases de dados. As consultas foram realizadas nos meses de março e abril do ano de 2020. Na construção da query de busca, em um primeiro momento, não foi colocada nenhuma restrição temporal.

Os descritores selecionados como argumentos para a busca, tanto em língua portuguesa, quanto em língua inglesa, foram: "Publicação de dados"; "Data publication"; "Data Publishing"; "Compartilhamento de dados"; "Data sharing"; "Artigo de dados"; "Data paper"; "Periódico de dados"; "Data Journal"; "Dados científicos"; "Scientific data", verificados no título, resumo e palavras-chaves.

Analisando-se o resultado da busca inicial, foi observado que a quantidade de referências retornadas anteriores ao ano 2000 que versavam sobre a Publicação de dados eram praticamente nulas. Desta forma, refinou-se a *query* de busca estabelecendo como novo critério temporal a busca por referências datadas a partir do ano 2000. Os resumos dos artigos recuperados foram lidos, de forma que foram selecionados para leitura completa e, posteriormente para análise e discussão, os artigos que abordavam temáticas relacionadas à Publicação de dados.

A análise e discussão dos artigos da pesquisa tomou como base o modelo de Publicação de dados desenvolvido por Kratz e Strasser (2014), devido ao seu potencial para tornar o processo mais claro e simples para os pesquisadores. Também pelo fato deste ter sido o único modelo focado em Publicação de dados encontrado nas pesquisas realizadas.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

A busca por definições (ou redefinições) em Ciência da Informação acerca de objetos sob seu tratamento é uma das tarefas fundamentais de seus pesquisadores (DIAS; VIEIRA; SILVA, 2013). Assim, ao se considerar a Publicação de dados de pesquisa, há a necessidade de definir, redefinir ou conceituar com maior precisão, o conceito de dados de pesquisa. A Organisation for Economic Co-operation and Development - OECD (Organização para a Cooperação e Desenvolvimento Econômico - OCDE) define dados de pesquisa como

> […] registros factuais (resultados numéricos, registros textuais, imagens e sons) usados como fontes primárias para a pesquisa científica e que são comumente aceitos na comunidade científica como necessários para validar os resultados da pesquisa. Um conjunto de dados de pesquisa constitui uma representação parcial e sistemática do sujeito que está sendo investigado (OECD, 2007, p.13, tradução nossa) [2].

Borgman (2015, tradução nossa, p.28)[3] explica o conceito de dados em um contexto relacionado à pesquisa como sendo "[…] representações de observações, objetos ou outras entidades usadas como evidência de fenômenos para fins acadêmicos ou de pesquisa". Seguindo linha de raciocínio similar, Assante *et al.* (2016) entendem que os dados de pesquisa estão relacionados com uma gama variada de materiais produzidos ao longo de atividades de pesquisa.

Embora não seja objeto desse estudo, pois para isso seria necessário um estudo de terminologia, ressalta-se que, frequentemente, podem ser encontrados nos textos científicos na área da Ciência da Informação o uso das expressões "dados de pesquisa" e "dados científicos" empregadas como sinônimos.

Essas duas expressões apresentam sutilezas que podem levar a compreensões diversas, porém, no contexto deste trabalho, será utilizada, preferencialmente a expressão "dados de pesquisa", na forma como conceituada pela OECD (2007). Sendo utilizada a expressão "dados científicos" apenas por questão de coerência, quando a referência original utilizada assim fizer.

As recomendações para a disponibilização de dados na *Web* (LÓSCIO; BURLE; CALEGARI, 2017; GOFAIR, 2020; OPEN KNOWLEDGE FOUNDATION, 2020) utilizam indistintamente os termos disponibilização, abertura, compartilhamento e publicação de dados, ao se referir ao processo de tornar os dados de pesquisa passíveis de recuperação, disponíveis para acesso, uso e reúso. Porém, nesse trabalho, que tem foco no pesquisador e descreve um processo que requer rigor na sua execução, será utilizado o termo Publicação de dados.

O processo de publicar dados de pesquisa consiste em um conjunto de ações que agrega valor a um conjunto de dados, possibilitando que os mesmos sejam acreditados por uma comunidade específica e que tenham a possibilidade de serem universalmente acessíveis através de ambientes de redes. Este acesso pode ser ao conteúdo integral do conjunto dos dados ou, conforme o esquema de licenciamento utilizado, a um subconjunto dos dados ou, ainda, apenas aos seus metadados descritivos.

Callaghan *et al.* (2013) esclarecem que a Publicação formal de dados agrega serviços aos conjuntos de dados que vão além do simples fato de tê-los disponíveis em um website. Ela inclui verificações de natureza estritamente técnicas, como o tipo do formato dos dados e metadados usados, até considerações de natureza mais científica, tal como verificar se os dados a serem publicados efetivamente possuem significado científico.

Austin *et al.* [4] (2017, p. 82, tradução nossa) definem a Publicação de dados de pesquisa da seguinte forma:

[2] Texto original em inglês: "[…] 'research data' are defined as factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated.".

[3] Texto original em inglês: "[…] representations of observations, objects, or other entities used as evidence of phenomena for the purpose of research or scholarship.".

[4] Texto original em inglês: "Research data publishing is the release of research data, associated metadata, accompanying documentation, and software code (in cases where the raw data have been processed or manipulated) for re-use and analysis in such a manner that they can be discovered on the Web and referred to in a unique and persistent way. Data publishing occurs via dedicated data repositories and/or (data) journals which ensure that the published research objects are well documented, curated, archived for the long term, interoperable, citable, quality assured and discoverable – all aspects of data publishing that are important for future reuse of data by third party end-users.".

> A publicação de dados de pesquisa é a liberação de dados de pesquisa, metadados associados, documentação acompanhante e código de software (nos casos em que os dados brutos foram processados ou manipulados) para reúso e análise de forma que possam ser descobertos na Web e referenciados de forma única e persistente. A publicação de dados ocorre através de repositórios de dados dedicados e/ou periódicos (de dados) que garantem que os objetos de pesquisa publicados estejam bem documentados, geridos e preservados, arquivados a longo prazo, interoperáveis, citáveis, com qualidade garantida e encontráveis - todos os aspectos importantes da publicação de dados que são importantes para o reúso futuro dos dados por usuários finais terceiros.

Com relação a esta definição de Publicação de dados de pesquisa elaborada por Austin *et al.* (2017), concorda-se com o posicionamento de Dallmeier-Tiessen *et al.* (2017), ao afirmarem que ela é consistente com as etapas necessárias para que um objeto digital de pesquisa esteja em conformidade com os princípios da iniciativa FAIR (Findable, Accessible, Interoperable e Reusable) (WILKINSON *et al.* 2016).

Ou seja, os dados de pesquisa devem ser encontráveis, acessíveis, interoperáveis e reutilizáveis (DIAS, G. A.; ANJOS, R. L.; RODRIGUES, 2019). Desta forma, entende-se *a priori*, que os dados publicados devem estar em sintonia com os princípios preconizados pela iniciativa FAIR (GO FAIR, 2020).

Kratz e Strasser (2014) indicam que a comunidade acadêmica, em sua maioria, concorda que a Publicação de dados é composta por três propriedades fundamentais (Propriedades de publicação), conforme pode ser observado na figura 1, sendo elas: (1) os dados precisam estar disponíveis; (2) os dados devem estar documentados; (3) os dados podem ser citados.

Os autores fazem, ainda, questionamentos acerca de uma quarta propriedade referente a (4) validação dos dados. Indagações acerca de como e em qual medida um conjunto de dados deve ser validado são postas.

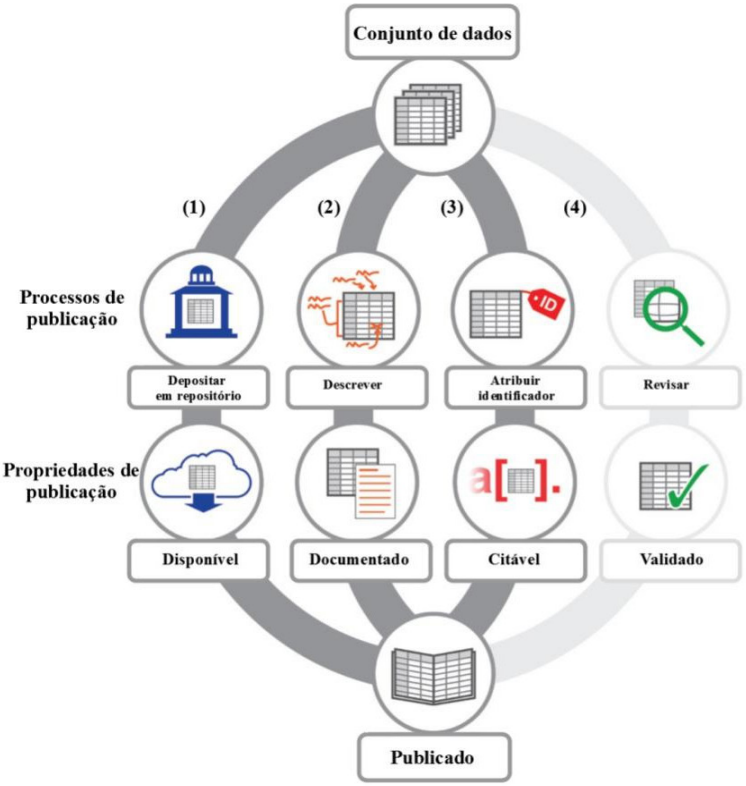
A um primeiro olhar, entende-se que a questão da validação de conjuntos de dados por terceiros no contexto de sua Publicação é uma ação que agrega valor a todo o processo, como será detalhado na subseção "Publicação de Conjuntos de Dados – Revisão".

Os processos de Publicação de dados apresentados por Kratz e Strasser (2014), também presentes na figura 1, estão estruturados em torno de ações que compreendem o (1) depósito de um conjunto de dados em um repositório; (2) sua descrição; (3) atribuição de um identificador; e um eventual (4) processo de revisão, o qual considera-se de fundamental importância. O trabalho de Kratz e Strasser (2014) sobre Publicação de dados pode ser compreendido como um modelo que pode ser ampliado e adaptado a casos concretos.

Alerta-se que, embora os Ciclos de Vida de Dados (CVDs) não sejam o objetivo desse trabalho, as ações que compreendem a Publicação de dados podem ser consideradas como subconjuntos das etapas existentes nos CVDs (ex: Modelo de Ciclo de Vida do DCC, DataONE, DDI, etc). A pesquisa de Araújo *et al.* (2019) apresenta uma compilação das etapas de três CVDs e pode ser utilizada para fazer um paralelo com as ações do modelo apresentado por Kratz e Strasser (2014).

Na sequência são apresentadas as etapas relativas aos processos de Publicação, com base no trabalho de Kratz e Strasser (2014), apresentando-se uma discussão sobre cada uma destas e alguns elementos de seu modelo.

Figura 1 – Publicação de conjuntos de dados



Fonte: Traduzido e adaptado de Kratz e Strasser (2014, p. 3).

PUBLICAÇÃO DE CONJUNTOS DE DADOS – DEPÓSITO

O depósito de um conjunto de dados constitui-se no processo de armazenar esses recursos em um repositório, preferencialmente em um repositório especificamente projetado para o depósito de dados de pesquisa. A disponibilização do conjunto de dados para a comunidade será efetivada após esta ação. Contudo, o depósito dos conjuntos de dados em um repositório de dados não implica que os mesmos estejam automaticamente disponíveis para qualquer pessoa. Estes dados podem estar protegidos por questões de cunho legal, como no caso de conteúdos acerca da saúde de pessoas, ou mesmo estar embargados por um período de tempo pelos seus criadores (KRATZ; STRASSER, 2014). Porém, mesmo no caso em que os conjuntos de dados não estão disponíveis, espera-se que um mínimo de informações estejam disponibilizadas acerca do seu conteúdo por meio do uso de metadados descritivos.

Os repositórios de dados científicos podem ser entendidos como instrumentos que proveem suporte para a Publicação de dados, disponibilizando recursos para todos os atores envolvidos no processo (ASSANTE *et al.*, 2016). Com relação a esses repositórios, Marcial e Hemminger (2010) esclarecem que os mesmos são conceitualmente similares aos repositórios institucionais, embora possuam naturezas muito diferentes no que tange a especificidades de domínio e em alguma medida no que diz respeito à utilidade.

No contexto da Publicação de dados, Curty e Aventurier (2017, p. 6, grifo nosso) trazem o seguinte comentário: "Os repositórios de dados são parte essencial do ecossistema da publicação de dados, e constituem por si só como uma abordagem de data publishing, uma vez que tornam públicas coleções de dados acompanhadas por recursos que otimizem seu potencial de reúso.".

Exemplos de produtos de software que podem ser utilizados para a construção de repositórios de dados incluem o Dataverse[5] e o DSpace[6]. Como implementações efetivas de serviços para o depósito de dados de pesquisa podem ser indicados o Dryad[7] , o Figshare[8] e o Zenodo[9] (ASSANTE *et al.*, 2016). No Brasil, ainda são poucas as instituições de ensino e pesquisa que disponibilizam serviços para o depósito de dados de pesquisa, foi detectado em uma investigação preliminar, ainda em andamento, que nem 20% das universidades federais possuem repositórios de dados.

Ressalta-se que na escolha do software, como afirma Torino, Roa-Martínez e Vidotti (2020), deve ser dada prioridade aos repositórios confiáveis, que possibilitem a preservação digital, a fim de garantir o acesso, uso e reúso dos dados a longo prazo. E, também, conforme o pensamento de Lóscio, Burle e Calegari (2017), que o repositório permita a utilização de padrões de metadados internacionalmente aceitos, a fim de possibilitar interoperabilidade e que os dados possam ser facilmente encontrados.

PUBLICAÇÃO DE CONJUNTOS DE DADOS – DESCRIÇÃO

O processo de descrição consiste em agregar informações adicionais aos conjuntos de dados, de modo a ampliar as possibilidades de interpretações sintáticas e semânticas destes objetos, tanto por seres humanos, quanto por sistemas automatizados, sendo este conjunto de informações que representa a documentação dos conjuntos de dados. Pode-se dizer que essas informações dão o contexto que proporcionará um maior e melhor entendimento sobre o conjunto de dados.

Existem vários tipos de insumos que podem contribuir na composição da documentação de um conjunto de dados. Neste sentido, Kratz e Strasser (2014) elencam os seguintes recursos: os metadados, o artigo cientifico tradicional e o artigo de dados. A estes recursos,sugere-se a inclusão, ainda, de ontologias e os Planos de Gestão de Dados (PGDs).

Ainda, no que diz respeito à elaboração dos metadados, indica-se que estes devem estar em sintonia com os princípios da iniciativa FAIR, contribuindo assim, para que sistemas computacionais possam achar, acessar, interoperar e reusar conjuntos de dados com um menor grau de intervenção humana (GOFAIR, 2020).

Um recurso que pode integrar a documentação de conjuntos de dados e que se entende ser essencial no processo de sua descrição são os artigos de dados. Chavan e Penev[10] (2011, p. 3, tradução nossa) explicam o que vem a ser um artigo de dados:

> Um artigo de dados é uma publicação de um periódico cujo objetivo é descrever dados, ao invés de relatar uma investigação. Como tal, contém fatos sobre dados, não hipóteses e argumentos em apoio a essas hipóteses baseadas em dados, conforme encontrado em um artigo de pesquisa convencional.

O que é complementado pela definição dada por Torino, Roa-Martínez e Vidotti (2020, p. 188), que afirma que um artigo de dados é "um objeto científico que descreve minuciosamente todos os elementos necessários para a compreensão do conjunto de dados, incluindo a justificativa e os métodos de coleta". Este objeto pode ser publicado em um periódico convencional ou em um periódico de dados (data jornal).

[5] Disponível para download a partir de https://dataverse.org/
[6] Disponível para download a partir de https://duraspace.org/dspace/
[7] Disponível a partir de https://datadryad.org/
[8] Disponível a partir de https://figshare.com/
[9] Disponível a partir de https://zenodo.org/

[10] "A data paper is a journal publication whose primary purpose is to describe data, rather than to report a research investigation. As such, it contains facts about data, not hypotheses and arguments in support of those hypotheses based on data, as found in a conventional research article.".

Os artigos de dados são importantes pelo fato de servirem como instrumentos que possibilitam um aumento na visibilidade dos dados, contribuindo, portanto, para o uso e reúso, e para que estes possam ser utilizados inúmeras vezes, inclusive para objetivos distintos, pois um cientista pode compreender mais facilmente o que um determinado conjunto de dados representa pela leitura de um artigo de dados associado, uma vez que ele conterá informações que vão desde os formatos de dados e sua localização, até a modalidade de política autoral adotada pelos detentores (SANT'ANA, 2019).

Assim, os artigos de dados além de contribuir para que dados sejam mais visíveis e passíveis de uso, reúso e compartilhamento, servem para divulgação da modalidade de política de direito autoral adotada pelo mesmo (SOUSA; DIAS; SOUSA, 2020). De modo análogo ao artigo científico tradicional, o artigo de dados pode trazer prestígio e reconhecimento da comunidade científica para seus autores. Para um maior aprofundamento nas questões que permeiam os artigos de dados, indica-se a pesquisa de Roa-Martinez *et al.* (2017), que propõe a estrutura comum de um artigo de dados baseada em conjuntos de metadados.

PUBLICAÇÃO DE CONJUNTOS DE DADOS – ATRIBUIÇÃO DE IDENTIFICADOR

Equivalente ao que acontece com artigos científicos, o uso de identificadores também pode ser usado para reconhecer de forma única um conjunto de dados. A persistência de um identificador é uma característica fundamental, pois permite que os conjuntos de dados sejam acessados de forma inequívoca, possibilitando a recuperação destes a partir de um endereço padrão associado ao identificador. É necessário mencionar que sempre que a localização de um determinado conjunto de dados for alterada, esta mudança deve ser refletida no identificador que referencia o respectivo recurso. Neste diapasão, Lawrence *et al.* (2011) indicam que o uso de um identificador é uma solução possível para abordar o problema de permanência no meio eletrônico.

A partir do momento em que o pesquisador dispõe de um identificador para um conjunto de dados, torna-se possível a sua citação inequívoca. Citar um conjunto de dados no artigo científico que resultou do seu uso possibilita que terceiros acessem este conjunto de dados e tenham a oportunidade de reutilizá-los. Ressalta-se que a citação de um conjunto de dados confere reconhecimento aos seus respectivos criadores.

Um objeto identificador pode ser utilizado como mecanismo no controle dos direitos autorais de produção intelectual em meio eletrônico, por permitir o uso de um único código de identificação para objetos usados em redes digitais. A Lei de Direitos Autorais (BRASIL, 1998) não contempla em sua totalidade as diversidades de criação do intelecto humano que surgem no ambiente atual, a exemplo dos bens intelectuais advindos das tecnologias de informação e comunicação (SOUSA; DIAS, 2012). Diante disso, um identificador pode ser visto como uma alternativa que pode contribuir na proteção do direito autoral no que diz respeito ao conteúdo, resguardando desse modo todas as características e atributos inerentes aos direitos morais e patrimoniais previstos em lei.

O Digital Object Identifier (DOI) é um exemplo de identificador que pode ser usado no contexto dos conjuntos de dados. Serviços de repositórios de dados como Dryad, o Zenodo e outros atribuem um DOI para cada conjunto de dados neles depositados. A figura 2 ilustra um exemplo de uma referência para um conjunto de dados depositado no Zenodo que dispõe sobre o número de casos de COVID-19 na Irlanda. O fragmento de texto destacado na figura 2 corresponde (10.5281/zenodo.3754983) ao DOI associado ao conjunto de dados. A obtenção do DOI pode ser realizada através de uma agência de registro. Duas das agências mais conhecidas são a Crossref [11] e a DataCite[12] . Ambas as agências proveem identificadores, contudo elas possuem objetivos distintos.

[11] Disponível a partir de: https://www.crossref.org/

[12] Disponível a partir de: https://datacite.org/

Figura 2 − Exemplo de citação para conjunto de dados

> Moriarty, Frank. (2020). Number of cases of coronavirus disease (COVID-19) in Ireland (Version v2.0) [Data set]. Zenodo. http://doi.org/**10.5281/zenodo.3754983**

Fonte: Moriarty (2020, grifo nosso).

A Crossref tem foco em atribuir identificadores para artigos de periódicos, artigos de conferências, livros e outros materiais, enquanto que o DataCite, por estar envolvido em uma ecologia mais voltada para a questão dos dados de pesquisa, está mais centrado neste tipo de objeto (CROSSREF, 2020). Assim, torna-se relevante a existência de um identificador único no processo de Publicação de conjuntos de dados, para resguardar a atribuição de autoria.

PUBLICAÇÃO DE CONJUNTOS DE DADOS – REVISÃO

A revisão de conjuntos de dados para a sua validação e posterior Publicação é uma etapa importante e que pode conferir confiança no uso destes recursos por terceiros. O processo de validação de conjunto de dados poderia, em alguma medida, mimetizar o processo de revisão pelos pares (*peer-review*) utilizado na revisão de artigos científicos (ASSANTE, 2016). Não obstante entender esta revisão como uma etapa necessária, questionamentos e dúvidas emergem com relação de como seria efetivamente este processo de validação (CALLAGHAN, 2013; KRATZ; STRASSER, 2014; ASSANTE, 2016).

É necessário refletir em que medida um eventual processo de revisão por pares de conjuntos de dados seria diferente do processo de revisão de artigos científicos, de modo que várias considerações podem emergir. Considera-se perfeitamente factível que um determinado parecerista, ao revisar um artigo que verse sobre sua área de domínio conhecimento, faça-o sem maiores dificuldades. Porém, no caso de ser atribuído para um parecerista um conjunto de dados, mesmo que ele entenda dos fatos representados pelos dados, pode ser que não entenda dos aspectos técnicos associados aos mesmos. Estes aspectos incluem: tipos de metadados usados, formatos de dados e produtos de software relacionados ao processo de captura e tratamento dos dados. Salienta-se ainda a questão de sobrecarga do trabalho a que os pesquisadores estão submetidos.

O tempo disponível para a revisão de artigos científicos é consideravelmente reduzido e compete com várias outras atividades de pesquisa com a qual estão envolvidos, assim, torna-se custoso ter de acoplar ainda. Desta forma, como seria possível equalizar a agenda de trabalho para ainda alocar tempo direcionado para a revisão de conjuntos de dados?

Callaghan (2013) comenta que, no ano de 2010, o periódico intitulado Journal of Neuroscience suspendeu a adição de qualquer arquivo suplementar ao artigo principal, visto a percepção de que os revisores estavam "imersos" pelo volume de dados e outros arquivos para avaliar, além do texto do artigo principal. Ainda com relação ao processo de revisão, Callaghan (2013) fez uma série de questionamentos. O autor refletiu sobre o que especificamente o parecerista deveria revisar e questionou se o que deve ser avaliado são os dados brutos, os metadados, o artigo de dados ou todos os elementos mencionados. O autor complementa indagando sobre que tarefa efetivamente deveria ser solicitada para o revisor executar. Outro ponto a considerar é que, ao contrário de um artigo científico, que uma vez publicado através de um periódico, normalmente não está sujeito a modificações, os conjuntos de dados podem passar por múltiplas versões ao longo do seu ciclo de vida. Desta forma, é possível levantar algumas questões adicionais sobre o processo de revisão de dados: como seria a revisão de um conjunto de dados que possui múltiplas versões e quem seria o parecerista em cada uma das versões dos artigos?

Sugestões para alguns dos questionamentos aqui postos são discutidos nos trabalhos de Klump *et al.* (2006), Lawrence *et al.* (2011) e Parsons e Fox (2013).

Os referidos autores apresentam ideias relacionadas, por exemplo, considerar a data da versão dos dados, onde o processo de revisão pode inclusive considerar aspectos da qualidade dos dados mantidos, bem como de seus metadados associados, devendo-se cada versão ser utilizada com bastante cautela.

Os revisores poderiam também incluir comentários sobre os dados que resultariam em possibilidade de mudanças em ambos, o que pode exigir que tanto os dados, quanto os metadados, possuam sempre novas versões correspondentes, que seriam publicadas em vez dos dados e metadados originais. Porém, esse processo não apresenta simplicidade, pois conduz a uma necessidade de se ter que solicitar aos autores que respondam aos revisores por mudanças ocorridas tanto nos dados, quanto nos metadados, podendo até mesmo requerer que os ajustes realizados por um curador também possam ser atualizados em reflexo ao processo de revisão. Todo esse processo pode ter necessidades específicas ao se considerar a devida proteção nos dados. Além, caso exista necessidade, da atribuição de direitos autorais, fato este que pode considerar diferenças culturais ou diferentes paradigmas de produção, a exemplo não só da qualidade dos dados, mas também dos aspectos espacial e temporal. Adicionalmente, considere-se incluso no processo o aumento de trabalho do parecerista/revisor para acompanhar as mudanças e ajustes.

A situação relativa à Publicação de dados pode apresentar ainda barreiras informacionais quando do compartilhamento dos dados, necessitando de procedimentos bem definidos para evitar inconsistências e o consequente desencontro do respectivo conteúdo, caso não se controle a persistência e confiabilidade dos dados e metadados subjacentes.

CONSIDERAÇÕES FINAIS

A partir das análises realizadas fica evidente que o processo de Publicação de dados não é uma atividade simples, mas, ao mesmo tempo, configura-se como primordial para o uso e o reúso de dados de pesquisa.

O domínio das atividades associadas com cada etapa da Publicação de dados (depósito, descrição, atribuição de identificador e revisão) pode transcender os saberes de uma única área do conhecimento. Desta forma, torna-se patente a necessidade de ter-se uma equipe composta por especialistas de diversas áreas, para que se possa efetivamente abordar a questão da Publicação de dados de pesquisa. Pois é improvável que um pesquisador domine, ao mesmo tempo, questões que envolvam repositórios de dados, uso de metadados, atribuição de identificadores, revisão de conjuntos de dados e mais uma série de questões tecnológicas subjacentes ao processo. Dentre os profissionais que poderiam contribuir com esta iniciativa podemos indicar, em um primeiro momento, o pessoal vinculado com as tecnologias da informação e comunicação, bibliotecários e arquivistas. Esta lista não é exaustiva, outros profissionais podem ser incluídos no processo, a partir das demandas específicas de cada caso concreto. Um caminho para a efetivação do processo de Publicação de dados para a comunidade de pesquisa passa pelo apoio das instituições a que estes pesquisadores estão vinculados. Isso demanda o estabelecimento de políticas relacionadas com a questão de dados no âmbito institucional, passando também pelo apoio (financeiro, de infraestrutura, etc.) das organizações governamentais relacionados ao desenvolvimento da ciência e tecnologia.

O assunto tratado nessa pesquisa tem potencial para gerar diversas outras pesquisas. Uma possibilidade de investigação futura seria avaliar o grau de maturidade das organizações de pesquisa no que diz respeito às etapas associadas com o processo de Publicação de dados.

REFERÊNCIAS

ARAÚJO, D. G. *et al.* Contribuições para a gestão de dados científicos: análise comparativa entre modelos de ciclo de vida dos dados. *Liinc em Revista,* Rio de Janeiro, v. 15, n. 2, p. 32-51, nov. 2019. Disponível em: https://doi.org/10.18617/liinc.v15i2.4686. Acesso em: 21 abr. 2020.

ASSANTE, M. *et al.* Are scientific data repositories coping with research data publishing?. *Data Science Journal,* [*s. l.*], v. 15, n. 6, p. 79–83, 2016. Disponível em: http://dx.doi.org/10.5334/dsj-2016-006. Acesso em: 12 abr. 2020.

AUSTIN, C. *et al.* Key components of data publishing: using current best practices to develop a reference model for data publishing. *International Journal on Digital Libraries,* [*s. l.*], v. 18, n. 2, p. 77–92, 2017. Disponível em: http://search-ebscohost-com.ez15.periodicos.capes.gov.br/login.aspx?direct=true&db=lih&AN=122919195&lang=pt-br&site=ehost-live. Acesso em: 21 abr. 2020.

BORGMAN, C. L. *Big data, little data, no data:* scholarship in the networked world. Cambridge: The MIT Press, 2015.

BRASIL. Lei nº 9.610, de 19 de fevereiro de 1998. Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências. *Diário Oficial da União*, Brasília, DF, 20 fev. 1998. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/L9610.htm. Acesso em: 27 abr. 2020.

CALLAGHAN, S. *et al.* Making data a first class scientific output: data citation and publication by NERC's Environmental Data Centres. *International Journal of Digital Curation,* [*s. l.*], v. 7, n. 1, p. 107–113, 2012. Disponível em: https://doi.org/10.2218/ijdc.v7i1.218. Acesso em: 12 abr. 2020.

CALLAGHAN, S. *et al.* Processes and procedures for data publication: a case study in the geosciences. *International Journal of Digital Curation,* [*s. l.*], v. 8, n. 1, p. 193-203, 2013. Disponível em: https://doi.org/10.2218/ijdc.v8i1.253. Acesso em: 12 abr. 2020.

CARLSON, J. The use of life cycle models in developing and supporting data services. In: RAY, J. M. (org.). *Research Data Management*: practical strategies for information professionals. West Lafayette: Purdue University Press, 2014. p. 63–86.

CHAVAN, V.; PENEV, L. The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC bioinformatics,* [*s. l.*], v. 12, n, 15, 2011. Disponível em: https://doi.org/10.1186/1471-2105-12-S15-S2. Acesso em: 12 abr. 2020.

CORTI, L. *et al*. *Managing and sharing research data*: a guide to good practice. Los Angeles: Sage, 2014. 234 p.

COSTELLO, M. J. Motivating Online Publication of Data. *BioScience,* [*s. l.*], v. 59, n. 5, p. 418–427, 2009. Disponível em: http://dx.doi.org/10.1525/bio.2009.59.5.9. Acesso em: 10 abr. 2020.

CROSSREF. *The basics.* 2020. Disponível em: https://www.crossref.org/community/datacite/. Acesso em: 10 abr. 2020.

CURTY, R.; AVENTURIER, P. O paradigma da publicação de dados e suas diferentes abordagens. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO - ENANCIB, 18., 2017, Marília, *Anais* [...]. Marília, SP. Disponível em: https://hal.archives-ouvertes.fr/hal-01637987. Acesso em: 15 abr. 2020.

DALLMEIER-TIESSEN, S. *et al.* Connecting data publication to the research workflow: a Preliminary analysis. *International Journal of Digital Curation,* [*s. l.*], v. 12, n. 1, p. 88-105, 2017. Disponível em: https://doi.org/10.2218/ijdc.v12i1.533. Acesso em: 13 abr. 2020.

DATA DOCUMENTATION INITIATIVE (DDI). *Why Use DDI*. [*S. l.*], 2020. Disponível em: http://www.ddialliance.org/training/why-use-ddi. Acesso em: 13 abr. 2020.

DATA OBSERVATION NETWORK FOR EARTH (DataOne). *Data Life Cycle.* [*S.l.*], 2020. Disponível em: https://www.dataone.org/data-life-cycle. Acesso em: 13 abr. 2020.

DIAS, G. A.; ANJOS, R. L.; RODRIGUES, A. A. Os princípios FAIR: viabilizando o reúso de dados científicos. *In*: DIAS, A. D.; OLIVEIRA, B. M. J. F (Orgs.). *Dados científicos:* perspectivas e desafios. João Pessoa: Editora UFPB, 2019. p. 177-187.

DIAS, G. A.; ANJOS, R. L.; ARAUJO, D. G. A gestão dos dados de pesquisa no âmbito da comunidade dos pesquisadores vinculados aos programas de pós-graduação brasileiros na área da Ciência da Informação: desvendando as práticas e percepções associadas ao uso e reúso de dados. *Liinc em Revista*, Rio de Janeiro, v. 15, n. 2, p. 5-31, 2019. Disponível em: http://revista.ibict.br/liinc/article/view/4683. Acesso em: 21 abr. 2020.

DIAS, G. A.; VIEIRA, A. A. N.; SILVA, A. L. A. Em busca de uma definição para o livro eletrônico: o conteúdo informacional e o suporte físico como elementos indissociáveis. In *Encontro Nacional de Pesquisa em Ciência da Informação,* 2013, Florianópolis. Disponível em: http://eprints.rclis.org/20904/. Acesso em: 10 abr. 2020.

DIGITAL CURATION CENTRE (DCC). *DCC Curation Lifecycle Model.* 2018. Disponível em: https://www.dcc.ac.uk/guidance/curation-lifecycle-model .Acesso em: 12 abr. 2020.

DRAZEN, J. M. *et al.* The importance and the complexities of data sharing. *The New England Journal of Medicine,* [*s. l.*], v. 375, n. 12, p. 1182-1183, 2016. Disponível em: https://www-nejm-org.ez15.periodicos.capes.gov.br/doi/10.1056/NEJMe1611027. Acesso em: 28 abr. 2020.

GOFAIR. FAIR *principles*. Disponível em: https://www.go-fair.org/fair-principles/. Acesso em: 16 abr. 2020.

KLUMP, J. *et al.* Data publication in the open access initiative. *Data Science Journal,* [*s. l.*], v. 5, p. 79–83, 2006. Disponível em: http://doi.org/10.2481/dsj.5.79. Acesso em: 10 abr. 2020.

KOWALCZYK, S. T. *Digital Curation for Libraries and Archives*. Santa Barbara, California: Libraries Unlimited, 2018.

KRATZ, J.; STRASSER, C. Data publication consensus and controversies. *F1000Research*, [s. l.], v. 3, n. 94, p. 1-21, 2014. Disponível em: https://doi.org/10.12688/f1000research.3979.3. Acesso em: 9 abr. 2020.

LAWRENCE, B. *et al.* Citation and peer review of data: moving towards formal data publication. *International Journal of Digital Curation*, [s. l.], v. 6, n. 2, p. 4-37, 2011. Disponível em: http://www.ijdc.net/article/view/181/265. Acesso em: 10 abr. 2020.

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. (ed.). *Data on the web best practices*. 2017. Disponível em: https://www.w3.org/TR/dwbp/. Acesso em: 14 jan. 2020.

MORIARTY, F. Number of cases of coronavirus disease (COVID-19) in Ireland. *Zenodo*, [*s. l.*], v. 2, 2020. Disponível em: http://doi.org/10.5281/zenodo.3754983. Acesso em: 10 abr. 2020.

MARCIAL, L. H.; HEMMINGER, B.M. Scientific data repositories on the Web: An initial survey. *Journal of the American Society for Information Science and Technology,* [*s. l.*], v. 61, n. 10, p. 2029-2048, 2010. Disponível em: https://doi-org.ez15.periodicos.capes.gov.br/10.1002/asi.21339. Acesso em: 14 abr. 2020.

OECD. Organisation for Economic Co-Operation and Development. *OECD Principles and Guidelines for Access to Research Data from Public Funding*. OECD Publishing, Paris, 2007. Disponível em: https://www.oecd.org/sti/inno/38500813.pdf. Acesso em: 10 abr. 2020.

PARSONS, M. A.; FOX, P. A. Is data publication the right metaphor?. *Data Science Journal,* [*s. l.*], v. 12, p. WDS32–WDS46, 2013. Disponível em: https://doi.org/10.2481/dsj.WDS-042. Acesso em: 13 abr. 2020.

ROA-MARTINEZ, S. M. *et al.* Estructura propuesta del artículo de datos como publicación científica. *Revista Española de Documentación Científica,* [*s. l.*], v. 40, n. 1, p.1-12, 2017. Disponível em: http://dx.doi.org/10.3989/redc.2017.1.1375. Acesso em: 14 abr. 2020.

RDP BRASIL. Práticas e Percepções dos Pesquisadores Brasileiros. *Repositórios Piloto da Rede Nacional de Ensino e Pesquisa*, v. 2, 2019. Disponível em: https://hdl.handle.net/20.500.12401/4. Acesso em: 21abr. 2020.

RICHARDSON, R. J. *Pesquisa social*: métodos e técnicas. São Paulo: Atlas, 2017.

SANT'ANA, R. C. G. Campo informacional resultante da interação de ciclos de vida dos dados. In: DIAS, G. A.; OLIVEIRA, B. M. J. F. (Orgs.) *Dados Científicos:* perspectivas e desafios. 1. ed., UFPB: João Pessoa, 2019. p. 33-52.

SOUSA, R. P. M.; DIAS, G. A.; SOUSA, M. R. F. Análise sobre dados abertos e regulação autoral no contexto da editoria científica. In: SHINTAKU, M. *et al.* (Orgs). *Tópicos sobre dados abertos para editores científicos*. Botucatu, SP: ABEC, 2020. 240 p. DOI: 10.21452/978-85-93910-04-3. p. 119-135. Disponível em: https://www.abecbrasil.org.br/arquivos/Topicos_dados_abertos_editores_cientificos.pdf. Acesso em: 14 abr. 2020.

SOUSA, R. P. M.; DIAS, G. A. Digital Object Identifier: uma breve reflexão sobre sua contribuição para proteção do direito autoral de obras literárias no meio digital. *In*: ALBUQUERQUE, M. E. B. C. *et al.* (Orgs.). *Representação da Informação*: um universo multifacetado. João Pessoa: Editora da UFPB, 2012. p. 141-156.

TORINO, E.; ROA-MARTÍNEZ, S. M.; VIDOTTI, S. A. B. G. Dados de pesquisa: disponibilização ou publicação?. In: SHINTAKU, M.; SALES, L. F; COSTA, M. (org).T*ópicos sobre dados abertos para editores científicos.* Botucatu, SP: ABEC, 2020. p. 183-201. DOI: 10.21452/ 978-85-93910-04-3.cap15.

UK DATA SERVICE. *Research data lifecycle.* [*S. l.*], [2020]. Disponível em: https://www.ukdataservice.ac.uk/manage-data/lifecycle. Acesso em: 13 abr. 2020.

WILKINSON, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data.* [*s. l.*], 2016. Disponível em: https://doi.org/10.1038/sdata.2016.18. Acesso em: 13 abr. 2020.

Uso de Dicionário Semântico de Dados na anotação de modelos de dados dimensionais para geração de indicadores de desempenho

Marcello Peixoto Bax
Pós-Doutorado pela Rensselaer Polytechnic Institute (RPI) - Estados Unidos. Doutor em Informática, Anal. Sistemas e Tratamento de Sinal pela Université Montpellier 2 - Sciences et Techniques (UM2) - França. Professor da Universidade Federal de Minas Gerais (UFMG) - Belo Horizonte, MG - Brasil.
http://lattes.cnpq.br/1864473087690223
E-mail: bax.ufmg@gmail.com

Evaldo de Oliveira da Silva
Doutorando Gestão & Organização do Conhecimento pela Universidade Federal de Minas Gerais (UFMG) - Brasil. Mestre em Ciência da Computação pela Universidade de Federal de Viçosa (UFV) - Viçosa, MG – Brasil.
http://lattes.cnpq.br/7337125039379689
E-mail: evaldo.oliveira@gmail.com

Submetido em: 30/04/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021.

RESUMO

Key Performance Indicators (KPIs) são usados por organizações para avaliar o desempenho de suas atividades, apoiando a decisão. Com esses indicadores, elas reveem seus processos, buscando a sua melhoria contínua. Em modelagem de dados, os modelos dimensionais estruturam os dados agrupando-os em "fatos" e "dimensões". Os fatos são representados por campos numéricos que permitem gerar KPIs. É importante, contudo, seguir técnicas e boas práticas de anotação de dados com metadados que minimizem interpretações divergentes. O modo de anotar dados é ilustrado com a técnica "Dicionário Semântico de Dados" (SDD), que os associa a conceitos e tem potencial para apoiar a geração de KPIs, enriquecendo-os e formalizando-os com ontologias. Seguindo essa técnica, é apresentado um breve experimento que anota um modelo de dados para cálculos de KPIs usando SDDs. Como resultado, o potencial dos SDDs no contexto da geração de KPIs em organizações é examinado. Conclui-se que, além da integração semântica dos dados, outra contribuição é a estruturação formal (em lógica) dos indicadores em grafos de conhecimento fundamentados por ontologias. Finalmente, o experimento contribui para a curadoria dos dados, já que o SDD segue as boas práticas e os princípios FAIR.

Palavras-chave: Modelos Dimensionais. Indicadores de desempenho. KPI. Dicionário de Dados. Ontologia. Anotação semântica. FAIR.

Annotation of data for generation of performance indicators in organizations

ABSTRACT

Key performance indicators (KPIs) are used by organizations to assess their performance, supporting the decision. With these indicators, they review their processes seeking continuous improvement. Dimensional models structure data by grouping it into "facts" and "dimensions." The facts represent numeric fields that leverage the generation of KPIs. However, it is essential to follow data annotation techniques and recommended practices with metadata seeking to minimize divergent interpretations. In the context of the generation of KPIs, described how an annotation occurs according to the method "Semantic Data Dictionary" (SDD), which associates data with concepts to generate these indicators, enriching and formalizing them using ontologies. A "use case" (experiment) of data annotation of a dimensional model for KPI calculations is presented, based on SDDs. As a result, the experiment examines the potential of applying SDDs in the context of generating organizational performance indicators (KPIs). Besides the conceptual integration of the data, it is possible to consider another contribution, which is the formal structuring (in logic) of the KPIs in graphs of knowledge grounded on ontologies. Finally, this work contributes to data curation since the SDD follows acceptable modeling practices (FAIR principles).

Keywords*: Dimensional Data. Performance Indicators. KPI. Data Dictionary. Ontology. Semantic Annotation.*

Anotación de datos para generar indicadores de desempeño en organizaciones

RESUMEN

Las organizaciones utilizan los principales indicadores de desempeño (KPI) para evaluar su desempeño, respaldando la decisión. Con estos indicadores, revisan sus procesos buscando una mejora continua. Los modelos dimensionales estructuran los datos agrupándolos en "hechos" y "dimensiones". Los hechos representan campos numéricos que aprovechan la generación de KPI. Sin embargo, es esencial seguir las técnicas de anotación de datos y las mejores prácticas con metadatos para minimizar las interpretaciones divergentes. En el contexto de la generación de KPIs, describió cómo se produce una anotación según el método "Diccionario de datos semánticos" (SDD), que asocia datos con conceptos para generar estos indicadores, enriqueciéndolos y formalizándolos a través de ontologías. Se presenta un "caso de uso" (experimento) de anotación de datos de un modelo dimensional para los cálculos de KPI, basados en SDD. Como resultado, el experimento examina el potencial de aplicar los SDD en el contexto de la generación de indicadores de desempeño organizacional (KPI). Además de la integración conceptual de datos, es posible considerar otro aporte, que es la estructuración formal (en lógica) de KPIs en grafos de conocimiento basados en ontologías. Finalmente, este trabajo contribuye a la conservación de datos, ya que el SDD sigue prácticas de modelado aceptables (principios FAIR).

Palabras clave*: Modelos dimensionales. Indicadores de desempeño. KPI. Diccionario de datos. Ontología. Anotación semántica.*

1 – INTRODUÇÃO

Um indicador chave de desempenho (KPI ou *Key Performance Indicator*) é um valor que pode ser medido e que demonstra a eficácia da organização em alcançar resultados (PARMENTER, 2015). KPIs permitem avaliar o atingimento de metas e rever processos para melhoria contínua das atividades, criando a base analítica para a tomada de decisões que priorizam as ações avaliadas (empiricamente) como mais relevantes. Eles medem, a título de ilustração, receitas, lucros, preços e custos, atividades, qualidade ou satisfação. Gestores e executivos interpretam KPIs para decidirem com base empírica, científica. Exemplo comum de mensuração é o percentual de aderência da realização de atividades ao que foi planejado anteriormente. Os KPIs podem ser vistos também no meio acadêmico. De acordo com Kolar, Harrison e Gliksohn (2018), eles podem avaliar o grau do alcance de objetivos de instituições de ensino ou programas de pesquisa. Além disso, são insumos para gerenciar e monitorar o atingimento de objetivos e auxiliar no planejamento estratégico. Para Kimball e Ross (2013), a criação de KPIs deve ser disciplinada por boas práticas de nomeação de dados. Em casos em que o conjunto de dados (*datasets*) utilizado para gerar os cálculos não seja de compreensão trivial, nomes podem ser atribuídos segundo diferentes interpretações. Com isso, os KPIs acabam resultando de combinações de dados incompatíveis, comprometendo os valores e prejudicando a tomada de decisão. No âmbito da geração de KPIs, é crucial garantir a qualidade dos dados, e a Curadoria Digital propõe técnicas de descrição de dados com metadados que favorecem a qualidade, a preservação e facilitam a descoberta de novas informações e novos conhecimentos pelo reuso de dados (MEDEIROS, 2018). No entanto, somente a definição dos metadados não basta para extrair dados dos sistemas de informação existentes e compartilhar *datasets*. Dados usados para geração de KPIs vêm de sistemas e modelos de dados distintos e requerem informações adicionais para que seus significados sejam explicitados.

Wise *et al.* (2019) afirmam que a pesquisa e o desenvolvimento na indústria biofarmacêutica também estão se tornando cada vez mais orientados por dados. Os autores destacam os princípios FAIR (*Findable, Accessible, Interoperable, Reusable*), propostos por Wilkinson *et al.* (2016), para aumentar a eficiência e a eficácia da gestão de dados científicos, não somente na produção biofarmacêutica, mas em outras áreas da indústria. Projetos de *data warehouse* têm sido construídos para alavancar a produção (WISE *et al.*, 2018; VAUDANO, 2013). Porém, Wise *et al.* (2019) destacam a necessidade de utilizar procedimentos para anotar e gerenciar os dados nesses projetos também, visando a atender aos princípios FAIR.

Rashid *et al.* (2020) avaliaram a aderência dos SDDs aos princípios FAIR, comparando-os a outras abordagens de anotação de dados. Na comparação, o SDD recebeu a maior nota (detalhes na Seção 3). Geralmente acompanhando *datasets*, os Dicionários de Dados tradicionais (não semânticos) facilitam o gerenciamento de dados. Para formalizar a descrição dos dados com os metadados dos dicionários, ontologias podem ser aplicadas. Elas enriquecem semanticamente e formalizam logicamente o significado dos dados, evitando interpretações discrepantes.

Este artigo apresenta um "caso de uso" de anotação de dados de um modelo dimensional para cálculos de KPIs, baseado na abordagem de Dicionários Semânticos de Dados (*Semantic Data Dictionary*, SDD) proposta por Rashid *et al.* (2017). Os SDDs contribuem para a curadoria dos dados e estão alinhados aos princípios FAIR. Trata-se do relato de um experimento que procurou examinar o potencial de uso de SDDs no contexto da geração de indicadores de desempenho organizacional (KPIs). A abordagem emprega ontologias para oferecer uma descrição detalhada de *datasets* a ponto de ser automatizada e tratada por computador. Ela permite o enfoque na semântica dos dados, incluindo informações que possam prontamente ser processadas.

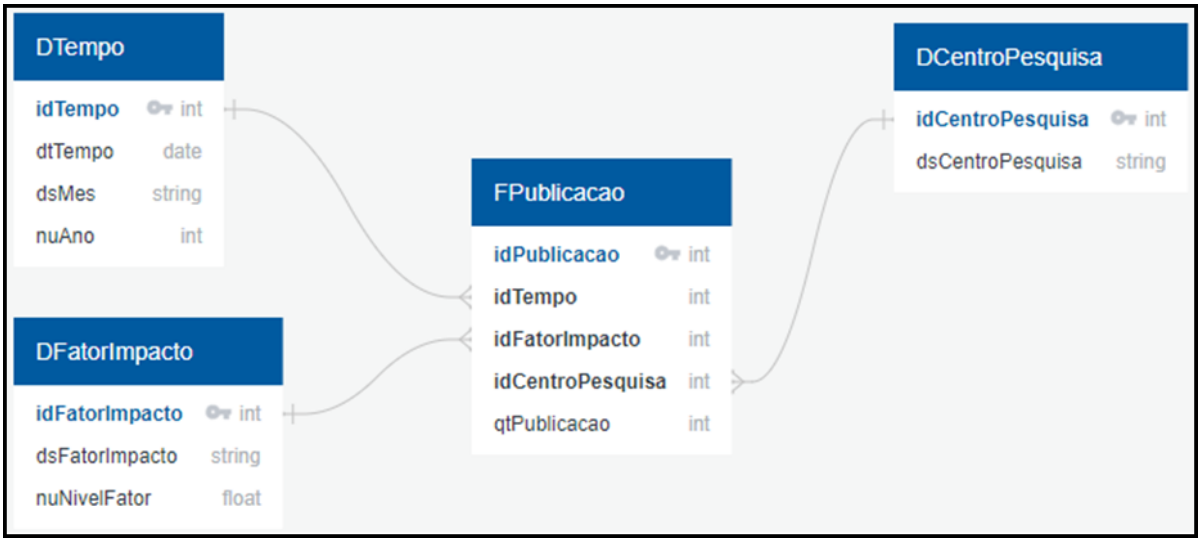
Ao considerar essas características dos SDDs, argumenta-se que é alcançada uma representação legível e padronizada por máquina para o registro de metadados com base em *datasets*, e em metodologias que usam linguagens de mapeamento. Isso é alcançado simplificando os requisitos de conhecimento de programação, separando a parte de anotação da abordagem do componente de software. Uma vantagem de aplicar SDDs em modelos dimensionais é alinhar e harmonizar interpretações de conceitos e diferentes escalas e unidades de medida que descrevem os dados. Pode-se assim integrar semanticamente dados de diferentes fontes e unidades de negócios ou até mesmo de organizações diferentes. Rashid *et al.* (2020) comparam a abordagem SDD com trabalhos correlatos por meio de métricas que possibilitam compreender as diferentes iniciativas de integração de dados. De acordo com os autores constatou-se desempenho superior da abordagem de SDD em relação, sobretudo, aos dicionários de dados tradicionais.

O texto está organizado como segue: a Seção 2 traz o conceito de modelagem dimensional de dados como KPIs e sua anotação por meio de ontologias. A Seção 3 descreve o processo de anotação usando o SDD. A Seção 4 aplica o SDD, relatando a anotação necessária para criar KPIs voltados ao monitoramento de quantitativos de publicações científicas em instituições de pesquisa. A Seção 5 analisa e relaciona os trabalhos correlatos de destaque na literatura. A Seção 6 traz as conclusões e considerações finais, além de sugerir trabalhos futuros.

2 – MODELANDO KPIS COM ONTOLOGIAS

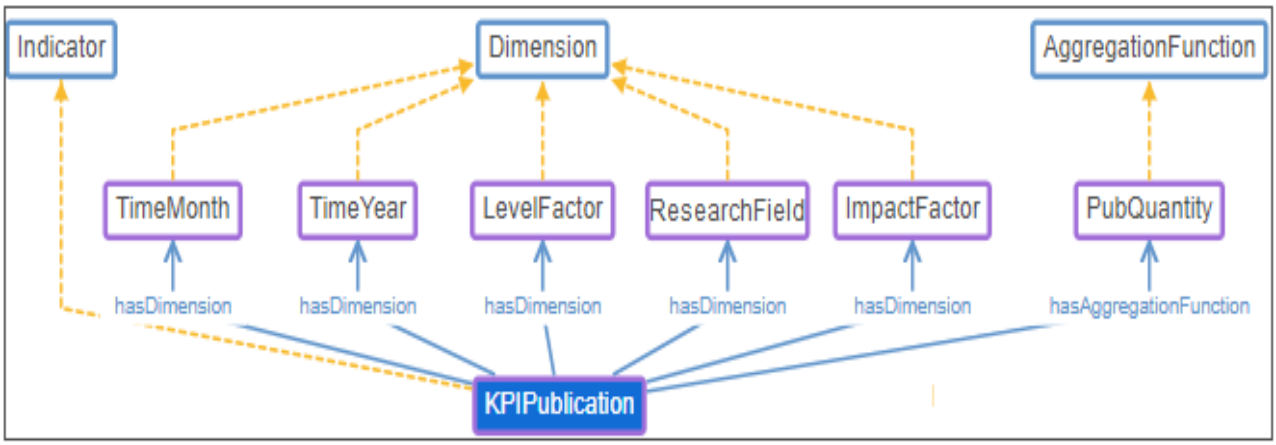
Um modelo de dados dimensional agrupa dimensões ou facetas ao redor de dados numéricos, chamados de fatos. A análise dos fatos usa as dimensões (ou facetas) para combinar filtros que atendem às necessidades do usuário e apoiam a tomada de decisão (KIMBAL; ROSS, 2013). O modelo dimensional da figura 1, organizado como uma estrela (fatos no centro), permite calcular os montantes de publicações por meio das seguintes dimensões: fator de impacto, centro de pesquisa, mês e ano. A tabela fato (FPublicacao) possui chaves (estrangeiras) oriundas das outras tabelas e um atributo quantidade de publicações (qtPublicacao), com o qual se extrai os KPIs a partir de funções de agregação e fórmulas.

Figura 1 – Exemplo de modelo dimensional de dados



Fonte: Elaborada pelos autores.

Figura 2 – Ontologia KPIOnto



Fonte: Adaptado pelos autores com base em Diamantini, Potena e Storti (2016).

Anotar dados semanticamente exige compreender o domínio representado por um modelo conceitual. Deve-se selecionar os dados e anotar os termos existentes que designam conceitos do domínio, explicitando e formalizando a sua semântica com o uso de ontologias. Esta anotação permite gerar fragmentos de conhecimento do domínio que podem ser representados por grafos de conhecimento (Hogan *et al.,* 2020). Um grafo de conhecimento representa objetos de interesse e conexões entre eles. Restrições na sua estrutura são impostas por meio de ontologias. Além disso, grafos de conhecimento permitem que pessoas e diferentes aplicações reutilizem as definições nele modeladas e gerem inferência de novos fatos, enriquecendo o conhecimento e o seu compartilhamento (Pan *et al.*, 2017).

Eles agregam entidades (e seus atributos), relacionando-as por relacionamentos expressivos e formais. Nessa perspectiva, eles são vistos como ontologias, nas quais as entidades formam o vocabulário que descreve os entes/indivíduos (instâncias) mais salientes e recorrentes do domínio.

O padrão RDF[1] (W3C, 2014) pode ser usado para representar grafos de conhecimento a partir de declarações no formato de triplas. Tais declarações (RDF) são geradas como fragmentos do conhecimento, como resultado da anotação por SDD. A conceitualização do domínio por meio da formalização semântica de ontologias é uma das premissas dos SDDs e este artigo utiliza a ontologia KPIOnto[2] de Diamantini, Potena e Storti (2016).

Ela nos permitirá anotar e alinhar conceitualmente a compreensão de diferentes profissionais sobre os KPIs empregados. A KPIOnto constitui-se de classes como: *Indicator, Dimension, AggregationFunction* e *Formula*; sendo "*Indicator*" a classe principal. Ela especifica um indicador pelas propriedades *hasDimension, hasFormula* e *hasAggrFunction* (para uso de funções de agregação). Como forma de representação das relações existentes na KPIOnto, a figura 2 apresenta instâncias que se relacionam com os conceitos sobre modelos dimensionais discutidos anteriormente. As instâncias foram criadas com o WebProtegé (MUSEN, 2015). Dessa forma, como mostra a figura 2, a instância *KPIPublication* (em azul) é um tipo de *Indicator*.

[1] *Resource Description Framework.*
[2] KPIOn to, https://github.com/KDMG/kpionto

A relação "*is-a*" é representada pela linha pontilhada amarela. A instância *KPIPublication* relaciona-se com os demais conceitos da KPIOnto, tais como *AggregationFunction* e *Dimension* pelas propriedades *hasAggrFunction* e *hasDimension*, respectivamente, representadas por linhas azuis sólidas. A instância *PubQuantity* é um de tipo de *AggregationFunction* pois representa o somatório da quantidade de publicações. As instâncias de *researchField, LevelFactor, ImpactFactor, TimeYear* e *TimeMonth* são um tipo de *Dimension*.

3 – ANOTANDO DADOS COM DICIONÁRIOS SEMÂNTICOS DE DADOS

Rashid *et al.* (2017) propõem o Dicionário Semântico de Dados (SDD) para anotar *datasets*. O SDD é uma abordagem de anotação de dados que emprega um conjunto de padrões de metadados fundamentados em ontologias que descrevem objetos (representados por dados) em classes e relacionamentos. A fim de enriquecer os dados presentes em um *dataset*, a anotação por SDD associa-os a conceitos/classes das ontologias. Recomenda-se o uso da ontologia *Semanticscience Integrated Ontology* (SIO) (2020), que fornece propriedades para descrever os relacionamentos entre objetos e atributos como modelo de representação do conhecimento e que permite também facilitar a descoberta do conhecimento. A abordagem deve ser orientada por especialistas de domínio, engenheiros de conhecimento (ontologistas) ou cientistas da informação que compreendem o domínio, tanto dos conceitos relacionados na ontologia, quanto dos *datasets* anotados. A anotação é um processo manual e utiliza um conjunto de documentos (templates de metadados) explicados mais abaixo no texto (*InfoSheet*; *Dictionary Mapping*; *CodeBook*; *Code Mapping*; *TimeLine*; *Properties Table*). A Seção 3 detalha esses templates.

A ferramenta *sdd2rdf*[3] (SEMANTIC DATA DICTIONARY, 2019) é um *script/software* que interpreta o SDD e converte os dados do *dataset* descrito por ele em um grafo de conhecimento expresso no padrão RDF.

[3] https://github.com/tetherless-world/SemanticDataDictionary

Para exemplificar o acesso aos dados anotados no grafo, o *sdd2rdf* cria alguns exemplos de consultas SPARQL[4]. O grafo de conhecimento (no formato RDF) gerado pelo script utiliza a ontologia e possibilita a interoperabilidade dos dados. É a formalização do vocabulário da anotação que abre caminho para interoperabilidade dos dados, que podem ser integrados de fontes diversas. Após escolher quais dados do *dataset* anotar, segue-se para a criação dos artefatos abaixo, em um processo cujas etapas são descritas a seguir:

- *Ontologia de domínio.* A ontologia formaliza os conceitos do problema de pesquisa. Deve-se buscar reutilizar ontologias consolidadas no domínio do problema;
- *Dictionary Mapping (DM).* Anota a semântica das colunas do *dataset*. Cada linha do DM mapeia uma coluna do *dataset*, formalizando-a conceitualmente, explicitando suas relações com os outros dados do mesmo *dataset*, bem como a sua proveniência[5] [6];
- *CodeBook*. Um *codebook* estrutura os dados categóricos[7] de um *dataset* mapeando-os para conceitos correspondentes na ontologia. Dessa forma, o cientista da informação preocupa-se com o tratamento dos dados, criando categorias e estabelecendo um código para cada uma. O *codebook* possui os seguintes campos para anotação: Coluna (entidade a ser anotada), Código, Descrição e Classe da Ontologia;

[4] *SPARQL Protocol and RDF Query Language* - Linguagem de consulta elaborada pelo W3C para acesso a dados em formato RDF.

[5] Proveniência de dados é a descrição das origens de um dado e o processo pelo qual ele chegou a uma base de dados (BUNEMAN; KHANNA; WANG-CHIEW, 2001).

[6] Como forma de anotar a proveniência do dado, o DM mapeia as entidades pré-existentes que são relevantes na anotação dos dados por meio do campo "*wasDerivedFrom*". Já o campo "*wasGeneratedBy*" descreve a atividade de geração associada à anotação de dados no DM (RASHID *et al*, 2020).

[7] Dados categóricos são dados agrupados. Podem derivar de observações feitas de dados qualitativos ou de observações de dados quantitativos agrupados em determinados intervalos (AGRESTI, 2003).

- *Infosheet*. Organiza os metadados de descrição do SDD. É importante principalmente para o seu compartilhamento em redes, conforme o princípio de "encontrabilidade" FAIR;
- *Grafo de Conhecimento (RDF).* Resulta da interpretação da dupla "SDD (templates de metadados) + Dados" pelo *script sdd2rdf*, gerando o grafo RDF. Caso seja necessário persistir os dados, o usuário pode armazenar o grafo em *triplestore*[8] para consulta posterior dos dados.

Inicialmente, os dados mapeados para as ontologias pelo SDD são as colunas do próprio *dataset*. Os objetos caracterizados nos *datasets* podem estar implicitamente representados. Os objetos implícitos serão explicitados no SDD e formalizados no grafo final gerado. A explicitação dos objetos implícitos favorece a integração semântica dos dados nos níveis conceituais mais abstratos do projeto, permitindo alinhar e homogeneizar, harmonizar, interpretações de conceitos que descrevem os dados que se deseja integrar.

Rashid *et al*. (2020) elaboraram métricas a fim de avaliar a aderência dos SDDs aos princípios FAIR. O SDD foi avaliado juntamente com outras abordagens de dicionário de dados, tais como: dicionários de dados tradicionais, abordagens envolvendo linguagens de mapeamento e ferramentas gerais de integração de dados. Para medir se o SDD atendeu a cada métrica, os autores forneceram um valor entre 0, 0,5 ou 1, dependendo do quanto o SDD responde a um parâmetro de avaliação. Em comparação às outras abordagens de dicionários de dados, o SDD recebeu nota 1.

Para pontuar a métrica localizável, foi avaliado o uso de identificadores persistentes exclusivos, como URLs, bem como a inclusão de metadados pesquisáveis, para que o conhecimento seja descoberto na web. O SDD permite a representação do conhecimento, podendo ser persistente e detectável. Para a métrica acessível, os autores consideraram que a representação de conhecimento gerada pelo SDD permite que os dados disponíveis estejam acessíveis abertamente, podendo ser divulgados publicamente. Em relação à métrica interoperável, os usos de vocabulários estruturados e de ontologias, como melhores práticas compatíveis com RDF, foram analisados. O SDD permite a representação do conhecimento a partir de vocabulários formais ou ontologias que são compatíveis com RDF. Sendo assim, para testar se o SDD permite reutilização dos dados, foi analisada a reutilização irrestrita deles. Também foi discutido se os metadados são detalhados o suficiente para um novo usuário entender. Constatou-se que o SDD permite o reuso irrestrito dos dados disponíveis publicamente.

4 – ANOTANDO DADOS PARA GERAR KPIS

Descreve-se, nesta seção, o exemplo de anotação de dados para geração de KPIs dada a necessidade de acompanhar índices quantitativos de publicação em centros de pesquisa. O modelo relacional da figura 1 foi utilizado como esquema de dados para recuperação dos *datasets*. Foram inseridos dados fictícios nas tabelas *DTempoMes, DTempoAno, DFatorImpacto, FPublicação, DCentroPesquisa*, onde o prefixo "D" é a designação de tabela de dimensão e o prefixo "F" é relativo à tabela fato para análise e geração de indicadores. O PostgreSQL foi usado como sistema gerenciador de banco de dados para persistir os *datasets*.

Para gerar o *dataset* a ser anotado, foi necessário relacionar as tabelas de dados (conf. figura 1). Esse estabelecimento de relação permitiu a criação de uma visão de dados (ou *view*) definida abaixo, gerando, como resultado, o *dataset* representado na figura 3 e que possui as colunas e dados correspondentes com a *view*.

[8] O armazenamento em *triplestore* (ou RDF) é um banco de dados com o propósito de armazenar e recuperar triplas por meio de consultas semânticas (DBPEDIA, 2020).

Figura 3 − *Dataset* a ser anotado

Id_Kpi	TimeMonth	TimeYear	ReasearchField	ImpactFactor	LevelFactor	PubWauantity
1	JANEIRO	2000	1	1	4	63
2	JANEIRO	2000	3	2	7	6

Fonte: Elaborada pelos autores.

Além da recuperação do *dataset*, é importante ressaltar que os elementos (*Dictionary Mapping*, *CodeBook* e *Infosheet*) utilizados em um SDD são definidos por meio de arquivos em formato CSV[9]. Abaixo segue a descrição da execução do processo de anotação:

- *Ontologia de Domínio*. As ontologias utilizadas foram a KPIOnto e a SIO. A KPIOnto possui conceitos consensuados que descrevem KPIs; já a SIO é a ontologia padrão utilizada nos SDDs;
- *Dictionary Mapping (DM)*. O DM (tabelas 1 e 2) mapeia, para ontologias (SIO e KPIOnto), as seguintes propriedades dos KPIs: dimensões *TimeMonth* (tempo na dimensão mês), *TimeYear* (tempo na dimensão ano), *ResearchField* (descrição do centro de pesquisa), *ImpactFactor* (descrição do fator de impacto) e *LevelFactor* (nível do fator de impacto); e função de agregação de *PubQuantity* (quantidade de publicação). Especificamente em relação à tabela 2, é possível identificar conceitos implícitos sobre o domínio. Dessa forma, os dados mapeados são de um *reseachInstitute*.Os dados implícitos são anotados no artefato DM do SDD para que sejam também enriquecidos semanticamente. Com isso, novos dados anotados aparecem, servindo de pontes para representar mais amplamente o conhecimento. É uma preparação para considerar novos dados na análise explicitando relacionamentos que até o momento estavam implícitos;
- *Codebook*. A tabela 3 traz o *Codebook*, que descreve os dados categoriais do *dataset,* mapeando-os para ontologias, nesse caso, a KPIOnto. São mapeadas as dimensões *DTempoMes, DTempoAno, DCentroPesquisa, DFatorImpacto, DNivelImpacto* e a função de agregação *QtPublicacao*;
- *Infosheet*. A tabela 4 possui os metadados (e seus vocabulários, *dct*[10], *owl*[11], *schema*[12]) que descrevem o SDD a fim de melhorar a sua "encontrabilidade" na rede (um princípio FAIR):
  - dct:creator: responsável pelo preenchimento;
  - dct:contributor: contribuidores na criação do *Infosheet* e execução do processo;dct:created: data de criação;
  - dct:description: propósito do SDD;
  - owl:imports: endereço das ontologias utilizadas no SDD;
  - schema:keywords: palavras-chave;
  - dct:publisher: responsável por publicar;
  - dct:title: título do SDD;
- *Grafo de Conhecimento*. O grafo RDF representando o conhecimento sobre os KPIs é persistido no Virtuoso[13], onde é possível manipular os dados empregando a linguagem SPARQL (ERLING; MIKHAILOV, 2009). Segue abaixo o trecho do grafo RDF (em sintaxe TTL[14]) que representa a anotação e integração semântica dos dados nas primeiras 4 linhas da tabela de dados da figura 3. Cada linha do *dataset* tem seus dados anotados pelos metadados do DM usando ontologias (tabelas 1 e 2).

[9] *Comma-Separated-Values*

[10] http://purl.org/dc/terms/
[11] https://www.w3.org/2002/07/owl
[12] http://schema.org/
[13] https://virtuoso.openlinksw.com/
[14] https://www.w3.org/TR/turtle/

Tabela 1 − Especificação do DM para objetos explícitos

Column	Attribute	sio:AttributeOf	rdfs:Label
Id_Kpi	sio:Identifier	??kpiPublication	Identificador do KPI
ResearchField	kpiOnto:hasDimension	??kpiPublication	Descrição do Centro de Pesquisa
ImpactFactor	kpiOnto:hasDimension	??kpiPublication	Descrição do Fator de Impacto
LevelFactor	kpiOnto:hasDimension	??kpiPublication	Nível do Fator de Impacto
TimeMonth	kpiOnto:hasDimension	??kpiPublication	Mês de Apuração do KPI
TimeYear	kpiOnto:hasDimension	??kpiPublication	Ano de Apuração do KPI
PubQuantity	kpiOnto:hasAggFunction	??kpiPublication	Quantidade de Publicação

Fonte: Elaborada pelos autores.

Tabela 2 − Especificação do DM para objetos implícitos

Column	**Entity**	**Relation**	**sio:InRelationTo**
??kpiPublication	kpiOnto:Indicator	kpiOnto:isUsedBy	??researchInstitute
??researchInstitute	sio:Institute	kpiOnto:hasKpi	??kpiPublication

Fonte: Elaborada pelos autores.

Tabela 3 − Codebook - dimensões DTempo, DFatorImpacto e DCentroPesquisa

Column	Code	Label	Class
DCentroPesquisa	1	CIÊNCIA DA INFORMAÇÃO	kpionto:researchField
DCentroPesquisa	2	CIÊNCIA DA COMPUTAÇÃO	kpionto:researchField
DCentroPesquisa	3	LINGUÍSTICA	kpionto:researchField
DFatorImpacto	4	IMPACTO ENTRE 2 E 4	kpionto:ImpactFactor
DFatorImpacto	5	IMPACTO ENTRE 5 E 7	kpionto:ImpactFactor
DFatorImpacto	6	IMPACTO ENTRE 7 E 10	kpionto:ImpactFactor
DNivelImpacto	7	4	kpionto:LevelFactor
DNivelImpacto	8	7	kpionto:LevelFactor
DNivelImpacto	9	5	kpionto:LevelFactor
DTempoMes	10	Janeiro	kpionto:TimeMonth
DTempoMes	11	Fevereiro	kpionto: TimeMonth
DTempoMes	12	Março	kpionto: TimeMonth
...			
DTempoAno	25	2000	kpionto: TimeYear
DTempoAno	26	2001	kpionto: TimeYear
DTempoAno	27	2002	kpionto: TimeYear
DTempoAno	28	2003	kpionto: TimeYear
DTempoAno	29	2004	kpionto: TimeYear
...			

Fonte: Elaborada pelos autores.

Tabela 4 − Especificação do Infosheet

Atributo	**Valor**
dct:creator	Marcello P. Bax e Evaldo de Oliveira da Silva
dct:contributor	Marcello P. Bax
dct:created	20/04/2019
dct:description	Anotação semântica do dicionário de dados para geração do KPI de Publicação
owl:imports	http://semanticscience.org/ontology/sio-subset-labels.owl
schema:keywords	KPI, Publicação
dct:publisher	Evaldo de Oliveira da Silva
dct:title	Geração de KPIs com base na anotação semântica de modelos de dados dimensionais

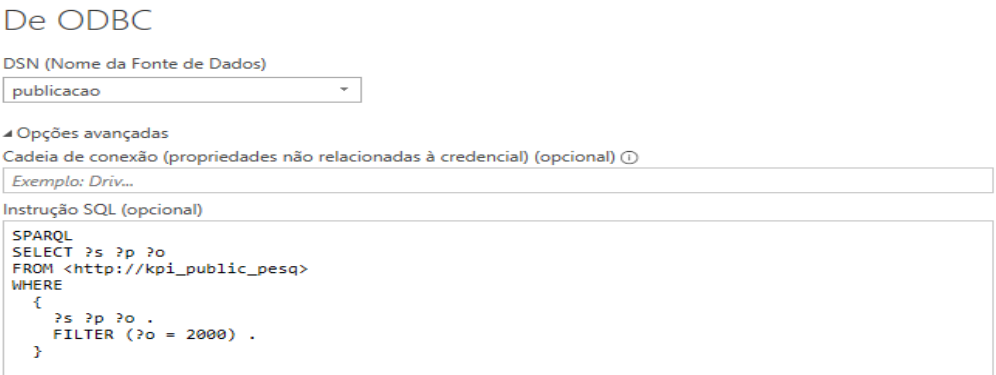
Fonte: Elaborada pelos autores.

```
example-kb:Id_Kpi-deb0 a example-kb:Id_Kpi , sio:Identifier ;
  sio:isAttributeOf example-kb:KpiPublication-ecc6 ;
  rdfs:label "Identificador do KPI"^^xsd:string ;
  sio:hasValue "1"^^xsd:integer .

example-kb:TimeMonth-50ac a example-kb:TimeMonth , kpiOnto:hasDimension ;
  sio:isAttributeOf example-kb:KpiPublication-ecc6 ;
  rdfs:label "Mes"^^xsd:string ;
  sio:hasValue "JANEIRO"^^xsd:string .

example-kb:TimeYear-65ea a example-kb:TimeYear , kpiOnto:hasDimension ;
  sio:isAttributeOf example-kb:KpiPublication-ecc6 ;
  rdfs:label "Ano"^^xsd:string ;
  sio:hasValue "2000"^^xsd:integer .

example-kb:ResearchField-896d a example-kb:ResearchField, kpiOnto:hasDimension ;
  sio:isAttributeOf example-kb:KpiPublication-ecc6 ;
  rdfs:label "Descricao do Centro de Pesquisa"^^xsd:string ;
  sio:hasValue "1"^^xsd:integer .

example-kb:ImpactFactor-3e96 a example-kb:ImpactFactor , kpiOnto:hasDimension ;
  sio:isAttributeOf example-kb:KpiPublication-ecc6 ;
  rdfs:label "Descricao do Fator de Impacto"^^xsd:string ;
  sio:hasValue "1"^^xsd:integer .

example-kb:LevelFactor-e2d7 a example-kb:LevelFactor , kpiOnto:hasDimension ;
  sio:isAttributeOf example-kb:KpiPublication-ecc6 ;
  rdfs:label "Nivel do Fator de Impacto"^^xsd:string ;
  sio:hasValue "4"^^xsd:integer .

example-kb:PubQuantity-8995 a example-kb:PubQuantity , kpiOnto:hasAggFunction ;
  sio:isAttributeOf example-kb:KpiPublication-ecc6 ;
  rdfs:label "Quantidade de Publicacao"^^xsd:string ;
  sio:hasValue "6"^^xsd:integer .

example-kb:Id_Kpi-f138 a example-kb:Id_Kpi , sio:Identifier ;
  sio:isAttributeOf example-kb:KpiPublication-e9c1 ;
  rdfs:label "Identificador do KPI"^^xsd:string ;
  sio:hasValue "2"^^xsd:integer .

example-kb:TimeMonth-50ac a example-kb:TimeMonth , kpiOnto:hasDimension ;
  sio:isAttributeOf example-kb:KpiPublication-e9c1 ;
  rdfs:label "Mes"^^xsd:string ;
  sio:hasValue "JANEIRO"^^xsd:string .

example-kb:TimeYear-65ea a example-kb:TimeYear , kpiOnto:hasDimension ;
  sio:isAttributeOf example-kb:KpiPublication-e9c1 ;
  rdfs:label "Ano"^^xsd:string ;
  sio:hasValue "2000"^^xsd:integer .

example-kb:ResearchField-08d4 a example-kb:ResearchField , kpiOnto:hasDimension ;
  sio:isAttributeOf example-kb:KpiPublication-e9c1 ;
  rdfs:label "Descricao do Centro de Pesquisa"^^xsd:string ;
  sio:hasValue "3"^^xsd:integer .

example-kb:ImpactFactor-ed7c a example-kb:ImpactFactor , kpiOnto:hasDimension ;
  sio:isAttributeOf example-kb:KpiPublication-e9c1 ;
  rdfs:label "Descricao do Fator de Impacto"^^xsd:string ;
  sio:hasValue "2"^^xsd:integer .
```

- *Visualização dos dados.* Um painel (dashboard) construído em Ms-PowerBI (MICROSOFT, 2020) conecta-se ao Virtuoso via ODBC (Open Database Connectivity) e executa consultas SPARQL para ilustrar como os dados, extraídos do grafo, podem ser visualizados de diferentes formas. A consulta SPARQL apresentada na Figura 4 é utilizada para extrair os dados do grafo de conhecimento, representando os dados sobre as publicações realizadas no ano 2000, para serem carregados para um arquivo do Ms-PowerBI. A partir da carga dos dados, deve ocorrer a sua "transformação" em formato RDF, para visualização de indicadores e métricas por meio dos dashboards.

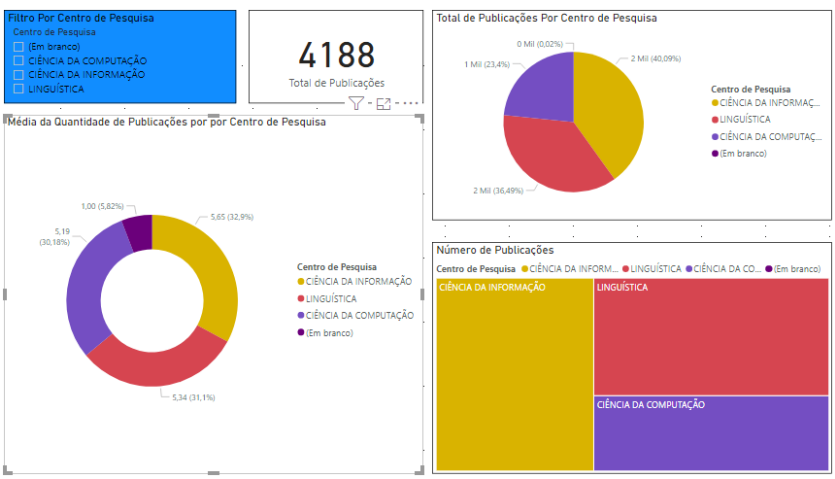
É importante ressaltar que é possível automatizar a carga e a transformação dos dados extraídos do Virtuoso quando o dashboard é criado no Ms-PowerBI, mesmo que novos dados sejam inseridos, alterados e excluídos nas fontes de origem. A figura 5 apresenta a visualização dos indicadores gerados, tais como o total geral de publicações, o percentual, a média e o número de publicações por centro de pesquisa. A figura 6 apresenta a visualização de outros indicadores, usando, como dimensão, o fator de impacto. Nesse caso, foram gerados os seguintes indicadores: mediana do número de publicações por fator de impacto e total de publicações por fator de impacto.

Figura 4 – Conexão via ODBC a partir de uma consulta SPARQL



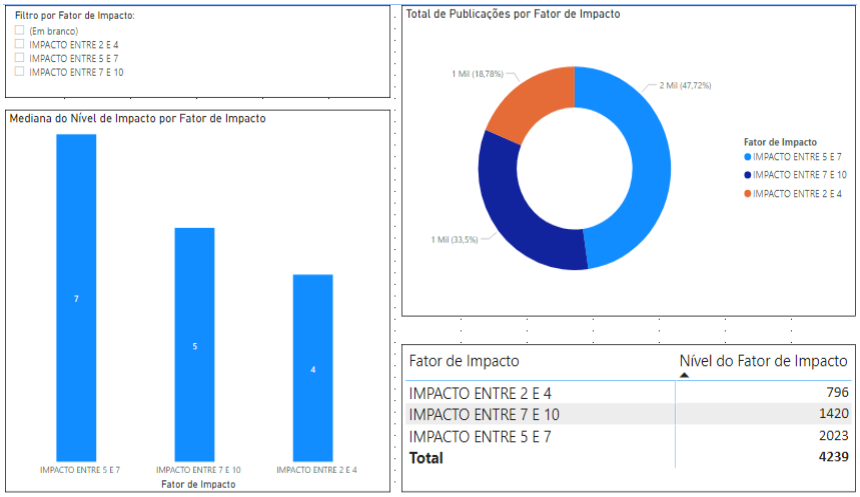
Fonte: Elaborada pelos autores.

Figura 5 – *Dashboard* gerado a partir dos RDFs com indicadores por Centro de Pesquisa



Fonte: Elaborada pelos autores.

Figura 6 - *Dashboard* gerado a partir dos RDFs com indicadores por Fator de Impacto



Fonte: Elaborada pelos autores.

5 – TRABALHOS CORRELATOS

Kritikos, Plexousakis e Woitsch (2017) afirmam que os dados conectados (*Linked Data*[15]) são um mecanismo para integrar fontes distintas, permitindo realizar inferências para derivar conhecimento novo. Eles utilizam essa ideia no contexto de negócios (BPaaS, *Business Process as a Service*), a fim de coletar e vincular informações originadas de diferentes sistemas. Propõem o uso de ontologias para melhorar a comparação de KPIs gerados dos dados integrados entre os sistemas. Wetzstein, Ma e Leymann (2008) indicam que KPIs sejam modelados por analistas de negócios que exploram anotações semânticas de processos de negócios. Os modelos de KPI são automaticamente calculados para serem geridos por meio de um painel de monitoramento em tempo real. Kourtesis, Alvarez-Rodrigues e Paraskakis (2014) sugerem uma estrutura semântica para gerenciamento de QoS (*Quality of Service*). Eles utilizam abordagens para o gerenciamento de QoS baseado em semântica, bem como os principais métodos e técnicas para explorar diversos dados. Silva *et al.* (2018) propõem um conjunto de funções para compor a estrutura semântica necessária à definição de dicionário de dados. Apresentam, ainda, como a estrutura semântica está relacionada à configuração sintática dos dicionários de dados, a fim de identificar padrões que possam ser usados no desenvolvimento de procedimentos para extração de informações e modelos semânticos.

O caso de uso apresentado neste artigo se diferencia dos trabalhos acima por utilizar a modelagem ontológica, aplicando a abordagem SDD para anotar dados de KPIs. A anotação é realizada manualmente, por especialistas de domínio. O resultado é a geração de grafos de conhecimento a partir de uma ontologia e de *templates* de metadados, sendo úteis para inferir novos conhecimentos, que podem melhorar a tomada de decisão dentro das organizações.

6 – CONSIDERAÇÕES FINAIS

No contexto organizacional, a modelagem conceitual adequada dos dados envolve a interpretação e negociação de significados sobre entidades, relacionamentos e regras de negócios, que ocorrem naturalmente na comunicação entre os vários atores (ou "partes interessadas"). As vantagens do SDD atingem pleno potencial em cenários onde a integração de fontes de dados diversas (internas ou externas) se faz necessária para enriquecer os dados que se quer analisar.

[15] O termo *Linked Data* refere-se a um conjunto de melhores práticas para publicar e conectar dados estruturados na Web (BIZER; HEATH; BERNERS-LEE, 2011).

O processo apresentado visa a organizar etapas para anotação com SDDs e geração do grafo de conhecimento (em RDF), representando formalmente o conjunto de fatos originados da combinação de dados de diferentes fontes. Um exemplo de geração de KPI, usando como fonte um modelo dimensional, para avaliar critérios de desempenho em função de publicações científicas produzidas por institutos de pesquisa ilustrou o processo, constituindo uma validação preliminar do método.

Argumentou-se que os SDDs contribuem para organizar e integrar dados oriundos de diferentes nichos da organização ou fora dela (por exemplo, via Web), gerando informações que estruturam conhecimentos sobre diversos indicadores empresariais (KPIs). Isso facilita os alinhamentos semânticos sobre os KPIs a partir de uma abordagem de modelagem de dados ampla, do tipo *top down*, e não apenas *bottom up*. O processo proposto permite associar dados de *datasets* a conceitos consensuados com a finalidade de gerar KPIs, enriquecendo-os e formalizando-os com ontologias. Além da simplificação da integração conceitual consensuada dos dados, outra contribuição é a estruturação formal (em lógica) dos KPIs em grafos de conhecimento fundamentados por ontologias.

Outras contribuições podem ser consideradas, como, a título de ilustração, para a comunidade da ontologia KPIOnto, que foi reutilizada no SDD. A abordagem descrita pode também contribuir para integrar fontes de dados heterogêneas, a exemplo de quando há necessidade de organizações diferentes trocarem informações sobre KPIs, como forma de monitoramento dos indicadores ou para atendimento a marcos regulatórios (estabelecidos por agências de regulação de diferentes mercados, como: energia, petróleo, saúde). Essa situação acontece frequentemente entre órgãos do governo, agências reguladoras e organizações empresariais, onde projetos de SDDs podem permitir o alinhamento conceitual de estados informacionais entre as organizações (por exemplo, sua situação financeira).

Nesse aspecto, o governo brasileiro já considera a dimensão semântica no desenvolvimento e na manutenção de ontologias, bem como em outros recursos de organização da informação, em vista de melhorar a interoperabilidade e a troca de informações, por meio do *e-Ping* (Interoperabilidade de Governo Eletrônico) (BRASIL, 2018). Finalmente, os SDDs contribuem para a curadoria dos dados, já que seguem boas práticas de modelagem (princípios FAIR).

Futuras pesquisas investigarão como a modelagem via SDD, tal como apresentada neste artigo, constituiria alternativa vantajosa à modelagem dimensional clássica (do tipo "*data mart*" ou "*data warehouse*"). Uma hipótese é que a flexibilidade de modelos conceituais ontológicos "livres de esquemas" (*schema free*) traria vantagens para a geração de KPIs no contexto analisado. Poderia, por exemplo, tornar a evolução do conhecimento sobre os indicadores de desempenho das organizações mais organizado, flexível, incremental e semanticamente enriquecido pela explicitação de sua semântica formal, advinda do uso de ontologias representadas em Lógica de Descrições (*Description Logic*) (KRÖTZSCH; SIMANCIK; HORROCKS, 2012).

REFERÊNCIAS

AGRESTI, A. *Categorical data analysis.* 2nd ed. New Jersey: Wiley, 2003.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data: The story so far. *In:* SHETH, A. *Semantic services, interoperability, and web applications*: emerging concepts. Hershey, USA: IGI Global, 2011. p. 205-227.

BRASIL. Programa de Governo Eletrônico Brasileiro. *e-PING Padrões de Interoperabilidade de Governo Eletrônico*. Brasil, Comitê Executivo de Governo Eletrônico, nov. 2018. Disponível em: http://eping.governoeletronico.gov.br/. Acesso em: mar. 2021.

BUNEMAN, P.; KHANNA, S.; WANG-CHIEW, T. Why and where: A characterization of data provenance. *In:* VAN DEN BUSSCHE, J.; VIANU V. (ed.). *Database Theory*: ICDT 2001. Berlin: Springer, 2001. p. 316-330. DOI: https://doi.org/10.1007/3-540-44503-X_20.

DBPEDIA. *About:* Triplestore. [*S.l.*], 2020. Disponível em: http://dbpedia.org/page/Triplestore. Acesso em: 16 set. 2020.

DIAMANTINI, C.; POTENA, D.; STORTI, E. SemPI: A semantic framework for the collaborative construction and maintenance of a shared dictionary of performance indicators. *Future Generation Computer Systems*, [*s.l.*], v. 54, p. 352-365, jan. 2016. DOI: https://doi.org/10.1016/j.future.2015.04.011. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0167739X1500103X. Acesso em: mar. 2021.

ERLING, O.; MIKHAILOV, I. RDF Support in the Virtuoso DBMS. *In:* PELLEGRINI, T.; AUER, S.; TOCHTERMANN, K.; SCHAFFERT, S (ed.). *Networked Knowledge-Networked Media:* integrating knowledge management, new media technologies and semantic systems. Berlin: Springer, 2009. p. 7-24.

FEW, S. *Information dashboard design:* The effective visual communication of data. [*S.l.*]: O'Reilly Media, 2006.

HOGAN, A. *et al.* Knowledge Graphs. *ArXiv preprint*, arXiv: 2003.02320, Mar. 2020. Disponível em: https://arxiv.org/abs/2003.02320. Acesso em: mar. 2021.

KIMBALL, R.; ROSS, M. *The data warehouse toolkit*: the definitive guide to dimensional modeling. 3rd ed. [*S.l.*]: Wiley, 2013.

KOLAR, J.; HARRISON, A.; GLIKSOHN, F. Key performance indicators of Research Infrastructures. *Central European Research Infrastructure Consortium*, Italy, ago. 2018. Disponível em: https://www.ceric-eric.eu/2018/08/30/key-performance-indicators-of-research-infrastructures/. Acesso em: 30 ago. 2018.

KOURTESIS, D.; ALVAREZ- RODRÍGUEZ, J. M. ; PARASKAKIS, I. Semantic-based QoS management in cloud systems: Current status and future challenges. *Future Generation Computer Systems*, [*s.l.*], v. 32, p. 307-323, 2014. DOI: https://doi.org/10.1016/j.future.2013.10.015. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0167739X1300232X. Acesso em: mar. 2021.

KRITIKOS, K.; PLEXOUSAKIS, D.; WOITSCH, R. Towards Semantic KPI Measurement. *In:* INTERNATIONAL CONFERENCE ON CLOUD COMPUTING AND SERVICES SCIENCE, 7., 2017, Portugal. *Proceedings* […]. Portugal: CLOSER, 2017. p. 91-102.

KRÖTZSCH, M.; SIMANCIK, F., HORROCKS, I. A description logic primer. *ArXiv preprint*, arXiv:1201.4089, jan. 2012. Disponível em: https://arxiv.org/abs/1201.4089. Acesso em: mar. 2021.

MEDEIROS, C. B. Gestão de Dados Científicos: da coleta à preservação. *SciELO em Perspectiva*, [*s.l.*], 22 jun. 2018. Disponível em: https://blog.scielo.org/blog /2018/06/22/gestao-de-dados-cientificos-da-coleta-a-preservacao/#.XXZ82ChKjIV. Acesso em: 4 set. 2019.

MICROSOFT. *PowerBI*. [*S.l.*], 2020. Disponível em: https://powerbi.microsoft.com/pt-br/. Acesso em: 24 abr. 2020.

MUSEN, M.A. *The Protégé project:* A look back and a look forward. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, jun. 2015. DOI: 10.1145/2557001.25757003. Disponível em: https://protege.stanford.edu/about.php. Acesso em: 23 abr. 2020.

PAN, J. Z. *et al.* (ed.). *Exploiting linked data and knowledge graphs in large organisations*. Switzerland: Springer, 2017.

PARMENTER, D. *Key Performance Indicators:* developing, implementing, and using winning KPIs. 3rd ed. New Jersey: Wiley, 2015.

RASHID, S. M. *et al.* The semantic data dictionary: an approach for describing and annotating data. *Data Intelligence*, v. 2, n. 4, p. 443-486, 2020. Disponível em: https://doi.org/10.1162/dint_a_00058. Acesso em: mar. 2021.

RASHID, S. M. *et al.* The semantic data dictionary approach to data Annotation and integration. *SemSci@ ISWC,* Vienna, Austria, p. 47-54, 2017. Disponível em: https://dblp.org/db/conf/semweb/semsci2017.html. Acesso em: mar. 2021.

SEMANTIC DATA DICTIONARY. *SDD Specification*. [*S.l*], 2019. Disponível em: https://github.com/tetherless-world/SemanticDataDictionary. Acesso em: 22 set. 2019.

SEMANTICSCIENCE INTEGRATED ONTOLOGY. *SIO.* [*S.l*], 2020. Disponível em: https://bioportal. bioontology.org/ontologies/SIO. Acesso em: 16 set. 2020.

SILVA, V. S.; HANDSCHUH, S.; FREITAS, A. Categorization of semantic roles for dictionary definitions. *ArXiv preprint*, arXiv:1806.07711, 2018. Disponível em: https://arxiv.org/abs/1806.07711. Acesso em: mar. 2021.

VAUDANO, E. The innovative medicines initiative: a public private partnership model to foster drug discovery. *Computational and structural Biotechnology journal*, [*s.l.*], v. 6, n. 7, p. e201303017, 2013. Disponível em: https://doi.org/10.5936/csbj.201303017. Acesso em: mar. 2021.

W3C. *RDF 1.1 Concepts and Abstract Syntax*. W3C Recommendation, Feb. 2014. Disponível em: https://www.w3.org/TR/rdf11-concepts/ . Acesso em 16 de set de 2020.

WETZSTEIN, B.; MA, Z.; LEYMANN, F. Towards measuring key performance indicators of semantic business processes. *In:* ABRAMOWICZ, W.; FENSEL, D. (ed.). *Business Information Systems*. Berlin: Springer, 2008. p. 227-238.

WILKINSON, M. D *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, [*s.l.*], v. 3, n. 160018, 2016. DOI: 10.1038/sdata.2016.18. Disponível em: https://www.nature.com/articles/sdata201618. Acesso em: mar. 2021.

WISE, J. *et al.* Implementation and relevance of FAIR data principles in biopharmaceutical R&D. *Drug discovery today*, [*s.l.*], v. 24, n. 4, p. 933-938, Apr. 2019. Disponível em: https://doi.org/10.1016/j.drudis.2019.01.008. Acesso em: mar. 2021.

WISE, J. *et al.* The positive impacts of real-world data on the challenges facing the evolution of biopharma. *Drug discovery today*, [*s.l.*], v. 23, n. 4, p. 788-801, Apr. 2018. Disponível em: https://doi.org/10.1016/j.drudis.2018.01.034. Acesso em: mar. 2021.

DBacademic: Conectando os dados abertos das instituições de ensino do Brasil

Sérgio Souza Costa
Doutor em Computação Aplicada pelo Instituto Nacional de Pesquisas Espaciais (INPE) - São José dos Campos, SP – Brasil. Professor da Universidade Federal do Maranhão (UFMA) - São Luís, MA - Brasil
http://lattes.cnpq.br/2073311645132958
E-mail: sergio.costa@ufma.br

Mateus Vitor Duarte Sousa
Graduando em Bacharelado Interdisciplinar em Ciência e Tecnologia pela Universidade Federal do Maranhão (UFMA) - São Luís, MA - Brasil
http://lattes.cnpq.br/7602019586262918
E-mail: mateusriograndense@gmail.com

Micael Lopes da Silva
Graduado em Engenharia da Computação pela Universidade Federal do Maranhão (UFMA) - São Luís, MA -Brasil.
http://lattes.cnpq.br/3876640219126946
E-mail: micaelopes32@gmail.com

Eddye Cândido de Oliveira
Graduando em Engenharia da Computação pela Universidade Federal do Maranhão (UFMA) - São Luís, MA - Brasil
http://lattes.cnpq.br/8610296132900395
E-mail: eddyeoliver@gmail.com

José Victor Meireles Guimarães
Graduando em Engenharia da Computação pela Universidade Federal do Maranhão (UFMA), São Luís, MA – Brasil.
http://lattes.cnpq.br/7974690550998329
E-mail: jvictormguimaraes@gmail.com

Submetido em: 24/11/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021.

RESUMO

As instituições públicas detêm um grande volume de dados que poderiam ser usados para melhorarem os seus serviços. Isso motivou um movimento denominado de dados abertos. Neste sentido, o Brasil e outros países, têm criado leis que incentivam e, de maneira compulsória, garantem que as instituições abram os seus dados públicos. Por meio do Decreto nº 8.777 de 2016 foi definido que todas as instituições federais deveriam elaborar o seu Plano de Dados Abertos (PDA). Esse decreto levou à criação e publicação de um grande volume de dados abertos pelas diversas instituições públicas. Atualmente, cada instituição mantém isoladamente os seus dados, o que torna impraticável a consulta de dados entre elas. O objetivo deste trabalho é, então, conectar esses dados em um grande repositório de dados denominado DBacademic. Para isso, dados abertos de 25 instituições públicas de ensino foram extraídos, transformados e carregados nesse repositório. Essa transformação resultou em quase 900 mil triplas que podem ser consultadas no endereço www.dbacademic.tech. Os resultados mostram o potencial dessa solução para possibilitar diversas consultas relevantes que seriam muito difíceis de serem realizadas com os repositórios isolados.

Palavras-chave: Dados Conectados. Universidades. RDF. PDA. Repositório.

DBacademic: Linking the open data of educational institutions in Brazil

ABSTRACT

Public institutions have a large amount of data that could be used to improve their services. This has motivated a movement called open data. In this sense, Brazil and other countries have created laws that encourage and, in a compulsory manner, guarantee that institutions open their public data. Through the Decree nº 8.777 of 2016, it was defined that all federal institutions prepared their Open Data Plan. This decree led to the creation and publication of a large volume of open data by the various public institutions. Currently, each institution maintains it's data in isolation, which makes it impossible to query data between them. The purpose of this work is then to connect these data to a large data repository called DBacademic. To that end, open data from 25 public educational institutions were extracted, transformed and uploaded to this repository. This transformation resulted in almost 900 thousand triples that can be queried through the following address: www.dbacademic.tech. The results showed the potential of this solution to enable several relevant queries that would be very difficult to be carried out as isolated repositories.

Keywords*: Open data. Universities. RDF. Open Data Plan. Repositories.*

DBacademic: Vinculando los datos abiertos de las instituciones educativas en Brasil

RESUMEN

Las instituciones públicas tienen una gran cantidad de datos que podrían utilizarse para mejorar sus servicios. Esto ha motivado un movimiento llamado datos abiertos. En este sentido, Brasil y otros países han creado leyes que fomentan y, de manera obligatoria, garantizan que las instituciones abran sus datos públicos. Mediante el Decreto nº 8.777 de 2016, se definió que todas las instituciones federales elaboraron su Plan de Datos Abiertos. Este decreto propició la creación y publicación de un gran volumen de datos abiertos por parte de las distintas instituciones públicas. Actualmente, cada institución mantiene sus datos de forma aislada, lo que hace que sea imposible consultar datos entre ellas. El propósito de este trabajo es entonces conectar estos datos a un gran repositorio de datos llamado DBacademic. Para ello, se extrajeron, transformaron y subieron a este repositorio datos abiertos de 25 instituciones educativas públicas. Esta transformación resultó en casi 900 mil triples que se pueden consultar a través de la siguiente dirección: www.dbacademic.tech. Los resultados mostraron el potencial de esta solución para permitir varias consultas relevantes que serían muy difíciles de realizar como repositorios aislados.

Palabras clave: *datos abiertos. Universidades. RDF. Plan de datos abiertos. Repositorios.*

1 – INTRODUÇÃO

No Brasil, o acesso aos dados de instituições públicas já estava previsto pela Constituição de 1988, porém foi reforçado por meio da Lei de Acesso à Informação (Lei n.º 12.527/2011). Adicionalmente, o Decreto 8.777, de maio de 2016, definiu que órgãos e entidades da administração pública federal deveriam elaborar e publicar um Plano de Dados Abertos (PDA). Em resposta a essa demanda, as instituições disponibilizam um grande volume de dados públicos e abertos em seus portais. Esses dados propiciaram uma maior participação da comunidade no desenvolvimento de soluções inovadoras que melhoram e fiscalizam os serviços públicos. Dentre essas instituições federais, este artigo focalizou as de ensino superior, técnico e tecnológico, como as Universidades Federais e os Institutos Federais de Ciência e Tecnologia (IFETs). Na literatura, é possível encontrar alguns trabalhos que destacam a demanda de acesso aos dados dessas instituições, como Carossi e Teixeira Filho (2017), Gama e Rodrigues (2016), Zorzal e Rodrigues (2016).

Muitas destas instituições de ensino possuem portais específicos para o acesso a seus dados abertos. Neste trabalho, foram identificadas 45 instituições, com uma média de 20 conjuntos de dados cada. Esses conjuntos de dados incluem informações sobre servidores, estudantes, projetos de pesquisa, despesas e orçamentos. Eles são indexados pelo Portal Brasileiro de Dados Abertos (http://www.dados.gov.br/) para facilitar a busca. Mesmo sendo indexados por este portal, tais dados permanecem isolados e de difícil integração. oAssim, diversas perguntas relevantes são difíceis de serem respondidas, como: Quais são os nomes (ou endereços eletrônicos) dos coordenadores dos cursos de Engenharia da Computação das instituições públicas de ensino do Brasil?

Mesmo com esses nomes presentes na maioria dos portais de dados abertos, esta pesquisa iria requerer um algoritmo específico para consultar cada um dos portais em busca dessa informação. Esse algoritmo teria que lidar ainda com os diferentes modelos de dados das instituições, tornando sua escrita impraticável, poi, a consulta manual nos portais Web seria provavelmente mais rápida e eficiente.

Esse é um exemplo simples, mas capaz de mostrar como a conexão entre os dados poderia aumentar muito a expressividade das consultas, o que facilitaria a análise de dados e o desenvolvimento de soluções inovadoras. Essa conectividade é possível a partir de um conjunto de boas práticas propostas por Tim Berners-Lee, denominadas Dados Conectados. Existem, atualmente, diversos repositórios de dados conectados, como pode ser observado no diagrama da figura 1. Nele, também é possível observar que o DBpedia é, provavelmente, o repositório mais conhecido e conectado (AUER *et al.*, 2007). No contexto de dados de instituições de ensino, alguns trabalhos têm proposto repositórios. Assim sendo, foram considerados, neste artigo, os estudos de Alencar *et al.* (2018), Costa *et al.*(2019), Kessler e Kauppinen (2015), Piedra *et al.*(2014), Rocha e Lóscio (2015) e Zablith, Fernandez e Rowe (2012).

Alencar et al (2018) aventaram a publicação e o consumo de dados conectados da Unidade Acadêmica de Informática (UAI) do IFPB (Instituto Federal da Paraíba) em conjunto com uma ontologia própria, denominada OpenUAI. Rocha e Lóscio (2015) também propuseram a publicação e a criação de uma ontologia para o Centro de Informática da Universidade Federal de Pernambuco (CIn/UFPE). Costa *et al.* (2019) apresentaram uma metodologia semiautomática para a extração de dados públicos e a sua transformação e carga como dados abertos conectados. Como caso de uso, os autores utilizaram os portais de dados públicos da Universidade Federal do Maranhão, ao invés do portal oficial de dados abertos. Em geral, esses artigos demonstram os ganhos de expressividade alcançados ao conectar os diversos dados de uma dada instituição de ensino.

Durante a pesquisa, muitos dos trabalhos citados não estavam em funcionamento. Além disso,aem sua maioria, eles não tinhammo objetivo de conectar os dados de toda a instituição. Outro desafio para integrá-los seria a necessidade de compatibilizar os seus distintos vocabulários.

Usando o conceito de dados conectados, este artigoovisa a propor um repositório de dados qur irá permitir responder perguntas que integrem dados de diferentes instituições de ensino do Brasil. Perguntas que atualmente seriam impossíveis, ou impraticáveis de serem respondidas. Este repositório, denominado DBAcademic, já agrega atualmente dados de 25 instituições Brasileiras de autarquia federal.

Este artigo está organizado da seguinte maneira: seção 2 apresenta alguns conceitos essenciais, seção 3 apresenta a metodologia do trabalho, seção 4 apresenta os principais resultados e desafios, e seção 5 apresenta as conclusões.

2 – FUNDAMENTOS

Existe uma linha tênue entre os conceitos de dados públicos, abertos e conectados. Nesse sentids, o conceito de dados é o mais básico e estudado pelos cientistas da informação. Semeler e Pinto (2019) apresentam muitas destas definições em um ensaio sobre os dados de pesquisa. Nesse ensaio, os autores consideraram os dados como qualquer objeto criado em formato digital, ou convertido para este, que possa ser usado para geração de *insights* de informação e conhecimento.

Os dados públicos são um subconjunto detses dados, que segundo a Controladoria Geral da União do Brasil, são informações que não lesem leis de privacidade, integridade e segurança (COSTA, *et al.*, 2013). No Brasil, a Lei nº 12.527, de novembro de 2011, regularizou o direito de acesso às informações de órgãos públicos administrativos, autarquias, fundações e empresas estatais (BRASIL, 2020).

Os dados abertos são aqueles que além de poderem ser usados, modificados e compartilhados precisam seguir alguns princípios, como: serem completos, primários, atuais, acessíveis, compreensíveis por máquina, não proprietários e livres de licença (OPEN GOVERNMENT WORKING GROUP, 2007; OPEN KNOWLEDGE FOUNDATION, 2019). Essa definição foi então reforçada pelo art. 2º do Decreto no 8.777/2016:

> III dados abertos-– dados acessíveis ao público, representados em meio digital, estruturados em formato aberto, processáveis por máquina, referenciados na internet e disponibilizados sob licença aberta que permita sua livre utilização, consumo ou cruzamento, limitando-se a creditar a autoria ou a fonte.

Os dados abertos atendem ao critério de transparência proativa, pois o detentor dos dados não espera uma solicitação para disponibilizá-los (ZORZAL; RODRIGUES, 2016). Etsa categoria de transparência deveria ser a forma principal seguida pelo estado para disponibilizar os seus dados, como destacado em (Zorza; e Rodrigues, 2016, p. 2)

> A informação sob a tutela do Estado é um bem público e sua evidenciação deve ser por iniciativa da Administração Pública, de forma espontânea, proativa, independente de qualquer solicitação, ou seja, transparência ativa, como definido em lei.

O Decreto no 8.777/2016 instituiu a política de dados abertos do poder executivo federal brasileiro e foi um importante passo nessa direção. Nesse decreto, definiu-se uma data limite para que órgãos e entidades da administração pública federal elaborassem e publicassem seus Planos de Dados Abertos (PDA). Atender a essa exigência foi, e ainda é, um grande desafio para muitas instituições federais, como destacado em Bertin *et al.* (2017) e Torino, Trevisan e Vidotti (2019). Mesmo com todos os desafios enfrentados pelas instituições, esse decreto ampliou muito a disponibilidade de portais de dados abertos das diversas instituições brasileiras.

Os dados abertos já têm um grande potencial para aumentar a participação da comunidade, melhorando a fiscalização e a qualidade dos serviços prestados pelas instituições públicas. Contudo, realizar consultas integrando estes dados é muito diiícil, sobretudo usando os formatos que são utilizados geralmente pelos portais de dados abertos. Bernes-Lee (2009) propõe então o conceito de dados conectados, chamando a comunidade para construir a "Web dos dados", em contraposto a atual "Web das páginas".

Em resumo, o conceito de dados conectados (ou ligados) refere-se a um conjunto de boas práticas para publicar e ligar os dados estruturados na Web (HEATH; BIZER, 2011). Dentre essas boas práticas, Tim Berners-Lee estabeleceu quatro princípios (BERNERS-LEE, 2009):

1. Use URIs (*Uniform Resource Identifier*) para identificar os recursos;
2. Use HTTP URIs de forma a possibilitar que as pessoas possam procurar esses recursos na Web;
3. Quando alguém procurar por uma URI, forneça informações relevantes utilizando formatos padrões;
4. Inclua conexões para outras URIs de forma a possibilitar que mais recursos possam ser descobertos.

O primeiro princípio exige o uso de *URIs* para identificar os recursos a serem publicadas,Eenquant, o segundo princípio assegura o uso desses identificadores por meio de requisições Web. Já o terceiro princípio estabelece que, quando solicitado um recurso, este deve ser fornecido com todas as suas informações em um formato de dados padrão mantido pela W3C[1], como o *Resource Description Framework* (RDF). Por último, o quarto princípio refere-se à conexão com dados já existentes.

Ao seguir esses princípios, será possível transformar a Web em um banco de dados global, como pode ser observado pelo diagrama criado pelo T*he Linked Open Data Cloud*[2] e ilustrado na figura 1. Nela é possível visualizar a conexão, representads por linhas, entre diversas bases de dados criadas e mantidas por diferentes instituições do mundo.

No centro da nuvem apresentada na figura 1, destaca-se a base de dados DBpedia. Esse projeto coletou os dados da Wikipédia e os disponibilizou no formato de dados conectados (MCCRAE *et al.,* 2020). Segundo McCrae *et al.* (2020), a DBpedia trata a Wikipédia como um banco de dados e tem o objetivo de extrair suas informações estruturadas e torná-las disponíveis na Web para qualquer outra base de dadosP permitindo, assim, incluir estes dados nas suas consultas, como será apresentado em um exemplo na seção 4.

Os dados abertos e conectados são aqueles que atendem a ambas as propriedades. Segundo (RIBEIRO 2015), os repositórios de dados abertos que seguem os princípios de dados conectados são hoje uma alternativa mais consistente e viável do que a disponibilização de documentos, arquivos com metadados. No esquema de classificação, proposto por Tim Berners-Lee (2009), os dados abertos e conectados possuem a avaliação mais alta em um sistema de classificação de 5 estrelas, figura 2.

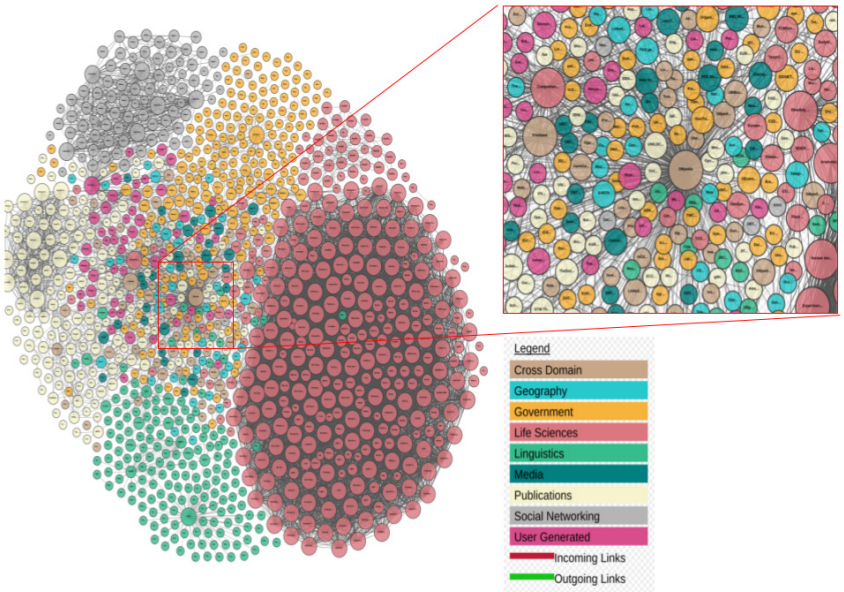
Esse sistema apresentado na figura 2 classifica o nível de abertura de dados. Assim, quanto mais alto o número de estrelas, mais fáceis esses dados estão de serem conectados. Cada estrela é atribuída a um nível de abertura dos dados, partindo da disponibilidade dos dados de maneira pública até chegar àqueles que estão conectados a outras bases. São eles:

1. Abastecimento dos dados públicos na Internet com licença aberta;
2. Utilização de formatos estruturados ao invés de páginas HTML;
3. Utilização de formatos não proprietários, como o CSV;
4. Emprego de padrões estabelecidos pela W3C, como o RDF;
5. Inclusão de conexões a outros dados já existentes.

[1] A W3C (www.w3.org) é a principal organização de padronização da World Wide Web.

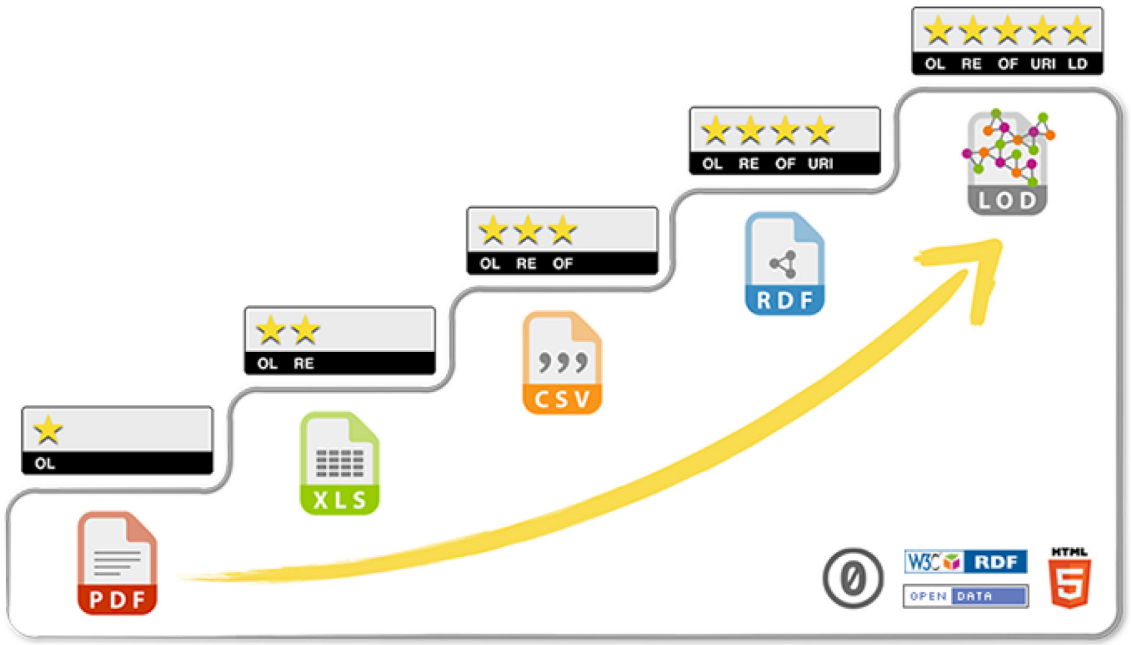
[2] O diagrama original pode ser acessado em https://lod-cloud.net/

Figura 1 – Diagrama de uma nuvem LOD em 2017



Fonte: Adaptado de McCrae *et al.* (2020).

Figura 2 – Classificação cinco estrelas



Fonte: Berners-Lee (2009).

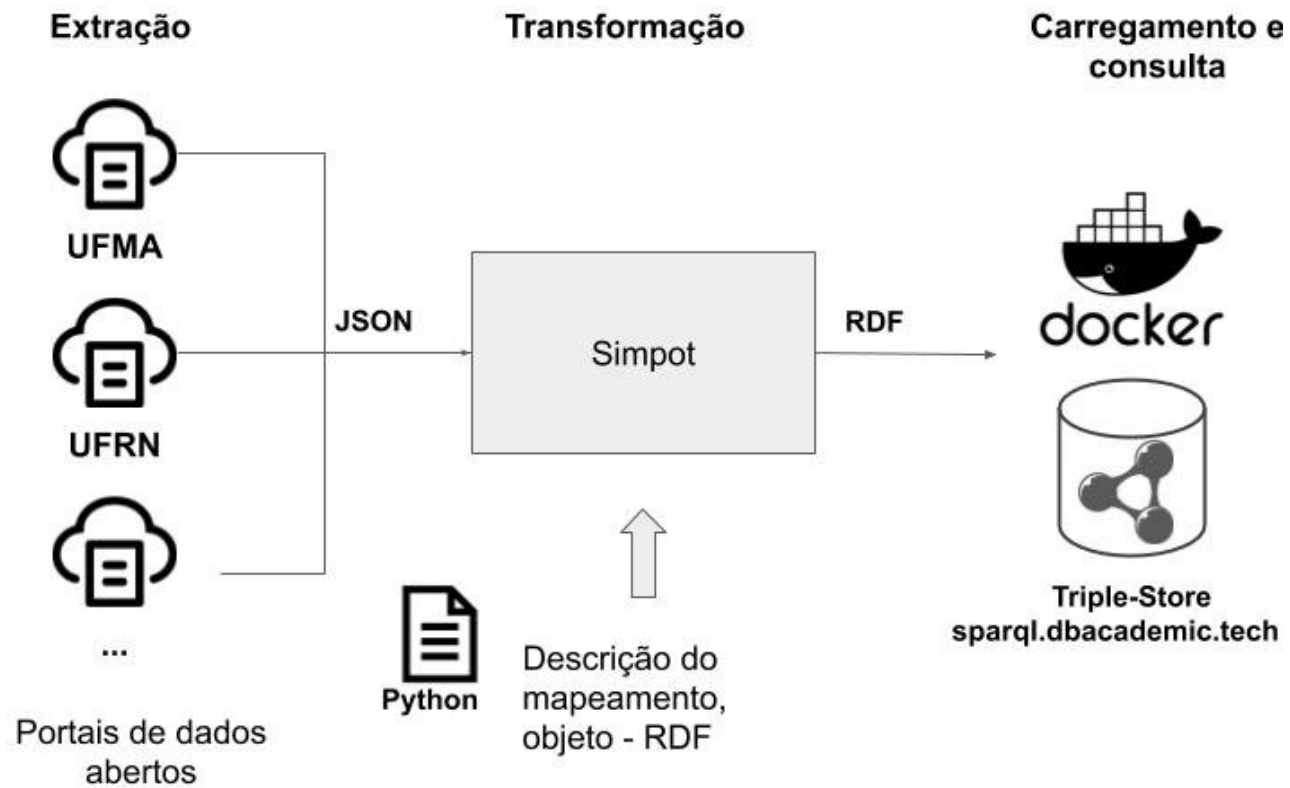
A primeira estrela é adquirida quando os dados são publicados sob licença aberta de uso. No entanto, características como a estruturação dos dados não sãé ainda algo fundamental. A segunda estrela é alcançada quando os dados já se encontram estruturados em formatos proprietários como o XLS da Microsoft. Adicionalmente, para receber a terceira estrela estes dados precisam estar em algum formato aberto, podendo ser CSV ou JSON. Para a próxima estrela já é necessário utilizar o formato de dados RD(. Esse formato de dados foi proposto pela W3C no contexto da Web Semântica e permite associar os recursos a conceitos de um ou mais vocabulários. Por fim, para atingir a quinta estrela, é necessário que os recursos se conectem, sempre que possível, a dados e vocabulários já existentes.

De acordo com esta classificação, os portais de dados abertos das instituições de ensino estariam classificados com 3 estrelas, enquanto o repositório DBacademic tem como objetivo atender todos os cinco princípios.

3 – METODOLOGIA

Para alcançar os objetivos propostos neste artigo, utilizou-se uma adaptação da abordagem semiautomática para a extração, a transformação e o carregamento de dados conectados proposto a em Costa *et al.* (2019). Essa metodologia foi influenciada pelo conceito *Extract Transform Load* (ETL), que é usualmente utilizado no contexto de Data Warehouse (VASSILIADIS; SIMITSIS; SKIADOPOULOS, 2002). Uma visão geral da metodologia é apresentada na figura 3.

Figura 3 − Visão geral da metodologia



Fonte: Elaboração do autor (2020).

Este trabalho difere, principalmente, em dois pontos àa abordagem proposta em Costa *et al.* (2019). oEm primeiro lugare os dados já são extraídos de portais de dados abertos ao invés de páginas Web. Em segundo lugare o objetivo do trabalho não está apenas na metodologia, mas na proposta de um repositório que irá conectar os dados acadêmicos de diversas instituições brasileiras, o DBAcademic. A seguir, os três passos apresentados na figura 3 serão detalhados.

3.1 EXTRAÇÃO DOS DADOS ABERTOS

Atendendo ao Decreto nº 8.777, de 11 maio de 2016, atualmente diversas instituições de ensino disponibilizam dados abertos. Essas instituições foram pesquisadasdno Portal Brasileiro de Dados Abertos, que indexa estes dados das instituições públicas.

Neste trabalho, identificou-se 45 instituições públicas de ensino com portais de dados abertos, sendo 25 universidades federais e 20 institutos federais. Entretanto, nessa primeira fase, foram incluídos os dados de 25 instituições que possuíam os conjuntos de dados mais acessíveis e completos. São elas: IFC, IFFAR, IFMA, IFMS, IFPA, IFPB, IFPI, IFRN, IFS, UFCA, UFCSPA, UFERSA, UFFS, UFMA, UFMS, UFOB, UFOP, UFPB, UFPEL, UFPI, UFRN, UFSJ, UFV, UNIFESSPA e UNIRIO.

Para a seleção dos dados, considerou-se os sete que eram os mais frequentes nos portais de dados abertos dessas instituições, como detalhado no quadro 1.

Quadro 1 − Descrições dos conjuntos de dados selecionados

Dados	Descrição
Docente	Informações de cada docente, como: nome, descrição, código, e-mail, áreas de interesses, departamento e URL para o currículo Lattes.
Curso	Informações de cada curso, como: nome, modalidade do curso, área de conhecimento do CNPQ, departamento, coordenador e título do profissional.
Departamento	Usualmente, os cursos e docentes são associados a um departamento. Desse modo, aqui incluem informações como: nome, localidade, chefe, centro no qual ele está associado e código.
Centro	Geralmente, o centro é uma unidade acadêmica em uma hierarquia superior aos departamentos. Os dados são, frequentemente: nome, localidade e diretor.
Grupo de Pesquisa	Os grupos de pesquisa são conjuntos de docentes e discentes que estudam um dado tema, tendo um docente como coordenador. Os dados são, usualmente: nome, área de conhecimento e coordenador.
Monografias (ou trabalhos de conclusão de curso)	Informações sobre as monografias (ou trabalhos de conclusão de curso) dos discentes, tais como: título, nome do aluno, nome do orientador, nome do curso, ano e data da defesa.
Discente	Este conjunto engloba as informações dos alunos ativos, ingressantes ou egressos da universidade. Geralmente, contém poucos atributos, como: nome, matrícula, período de ingresso e nome do curso.

Fonte: Elaborado pelo autor (2020).

Importante destacar aqui que, diferentemente de Costa *et al.* (2019), neste trabalho, não foram extraídos dados de páginas Web, mas sim de portais de dados abertos. Como foi discutido na seção 2, esses portais já disponibilizam dados em formatos abertos e acessíveis por algoritmos computacionais. Assim, para a extração, foi necessário apenas identificar os endereços desses recursos. Por exemplo, o recurso docente da Universidade Federal do Rio Grande do Norte é acessível através do seguinte endereço: http://dados.ufrn.br/api/action/datastore_search?resource_id=ff0a457e-76fa-4aca-ad99-48aebd7db070

Esses dados podem ser acessados por meio de navegadores Web, como o Chrome ou o Firefox. Contudo, para a automação, o ideal é escrever um código em alguma linguagem de programação que possa extrair e processar esses dados. Neste trabalho, foi utilizada a linguagem de programação Python[3].

3.2 TRANSFORMAÇÃO PARA DADOS CONECTADOS

Antes de descrever esse processo, é importante destacar que os dados conectados representam um paradigma diferente para a representação da informação. Nos portais de dados abertos, os dados são usualmente disponibilizados em formatos tabulares (linhas e colunas) ou como coleção de objetos com suas propriedades e valores. Os dados conectados são construídos a partir de três blocos básicos (ISOTANI; BITTENCOURT, 2015):

1. Modelo de dados padrão;
2. Vocabulários de referência;
3. Protocolo padrão de consulta.

Tanto o modelo de dados padrão quanto o vocabulário de referência utilizam, geralmente, o RDF que foi proposto e é mantido pela W3C para representação de metadados. Esse formato representa os dados como coleções de afirmativas (ou triplas) declaradas por um sujeito, um predicado e um objeto.

O sujeito e o objeto correspondem aos recursos a serem conectados, enquanto o predicado caracteriza a natureza dessa conexão direcionada do sujeito ao objeto. Um predicado também pode ser denominado de propriedade. Um objeto, em algum momento, pode ser um dado literal como um número, uma data ou um texto. Na seguinte tripla, por exemplo, o objeto é um texto, especificamente, o nome de Leonardo Da Vinci.

- **Sujeito:** http://dbpedia.org/page/Leonardo_da_Vinci
- **Predicado:** <http://dbpedia.org/ontology/birthName>
- **Objeto:** "Leonardo di ser Piero da Vinci"

Diferentemente do objeto, o sujeito precisa sempre ser um recurso e estar associado a um identificador único (URI). Neste trabalho, os sujeitos são recursos como docentes, discentes, cursos e departamentos. Sendo assim, é necessário definir um esquema para a criação de URIs para todos eles. Para isso, foi registrado um domínio onde cada recurso é associado ao seguinte esquema: www.dbacademic.tech/resource/<codigo único>. O código único foi gerado por meio do algoritmo de sintetização de mensagem MD5 (RIVEST, 1992). Em resumo, esse algoritmo retorna um código de 128 bits para um dado texto. Como entrada, usou-se um texto que é a concatenação (ou união) da sigla da instituição, o nome e um código do recurso. A sigla da universidade e o nome do recurso foram necessários para garantir um código único entre os recursos e as instituições.

Além dos sujeitos, os predicados também precisam estar associados a uma URI, que, nesse caso, representa um vocabulário. No exemplo anterior, o predicado birthName faz parte de um vocabulário criado pelo projeto DBpedia e está associado ao endereço: http://dbpedia.org/ontology/birthName. Muitos projetos precisam criar sua própria ontologia, que, no contexto computacional, são especificações formais e explícitas de conceitualizações compartilhadas e servem como base para garantir uma comunicação livre de ambiguidades (BREITMAN, 2006).

[3] Mais informações sobre essa linguagem podem ser encontradas em https://www.python.org/.

Além disso, segundo os quatro princípios dos dados conectados, eles devem, sempre que possível, se conectar a dados e vocabulários já existentes. Esse é um importante princípio, que permitiu construir uma base de dados global como a ilustrada pela figura 1. Desse modo, utilizou-se, como principal fonte de referência, as ontologias e os vocabulários citados nos estudos de Alencar *et al.* (2018), Costa *et al.* (2019); Kessler e Kauppinen, (2015), Piedra *et al.* (2014), Rocha e Lóscio (2015), Zablith, Fernandez e Rowe (2012) . Nesses trabalhos destacaram-se os vocabulários descritos no quadro 2:

Quadro 2 − Descrições dos vocabulários pesquisados

Vocabulário	Prefixo	Descrição
Academic Institution Internal Structure Ontology	AIISO	Descreve a estrutura organizacional interna de uma instituição acadêmica.
Dublin Core	DC	Descreve metadados genéricos.
Bibliographic Ontology Specification	BIBO	Descreve citações e referências bibliográficas.
Friend Of A Friend	FOAF	Descreve o vínculo de pessoas, seja por informações formais, como documentos físicos e digitais, seja por relacionamentos não formais.
Corresponde ao VCF (Virtual Contact File)	VCARD	Descreve pessoas e organizações utilizando *técnicas da web semântica.*
DBpedia Mappings Wiki	DBO	Descreve a semântica da extração dos dados da Wikipedia.
Open Information Center	OpenCIn	Descreve o ambiente acadêmico no qual o Centro de Informática da UFPE está imerso, essencialmente focalizado nos docentes e entidades relacionadas.
Web Ontology Language	OWL	Usado para representar e instanciar ontologias na World Wide Web.
Organization Ontology	ORG	Descreve estruturas organizacionais.

Fonte: Elaborado pelo autor (2020).

Para a seleção dos vocabulários, optou-se por aqueles mais consolidados e utilizados nos trabalhos anteriores. Contudo, como será descrito na seção 4, identificaram-se alguns desafios para o reuso de alguns deles. Dessa maneira, deverá ser proposta uma ontologia específica para as instituições de ensino em um trabalho futuro. Essa ontologia irá reusar muitos dos vocabulários apresentados no quadro 2, mas terá que incluir novos termos que representem melhor alguns conceitos das instituições de ensino. Além disso, espera-se que esse vocabulário possa ser construído por meio de parceria entre pesquisadores de diferentes instituições.

Após a definição de um esquema de geração de URIs e a seleção de vocabulários, é possível realizar a transformação entre os formatos de serialização. Estes são modelos de representação de dados que definem regras de sintaxe e esquemas para validação da informação. Os formatos de serialização mais comuns nos portais de dados abertos são: *JavaScript Object Notation* (JSON) e *Comma-Separated Values* (CSV). Em vista disso, foi necessária a transformação desses formatos em outro que fosse capaz de representar coleções de triplas, como o RDF/XML. Lembrando que o RDF é apenas um modelo de dados abstrato, enquanto o RDF/XML é um formato de serialização. A quadro 3 apresenta um fragmento de um documento RDF serializado como RDF/XML.

Quadro 3 − Representação de um docente por meio do RDF/XML

```
<rdf:Description rdf:about=”https://www.DBacademic.tech/resource/b39ba05094dd03dee6515349c07661a1”>

<foaf:name>LEVINDO DINIZ CARVALHO</foaf:name>

<owl:sameas rdf:resource=”https://sig.ufsj.edu.br/sigaa/public/docente/portal.jsf?siape=1943390”/>

<cin:siape>1943390</cin:siape>

<rdf:type rdf:resource=”http://dbpedia.org/ontology/Professor”/>

</rdf:Description>
```

Fonte: Elaborado pelo autor (2020).

A transformação foi realizada por intermédio da biblioteca *Simple Object-triple Mapping*[4] (SIMPOT), proposta em Costa *et al.* (2019). Uma vantagem dessa biblioteca é que ela permite fazer o mapeamento entre os modelos de objeto e tripla de forma declarativa. Essa abordagem tornou o processo de transformação mais fácil de replicar para os dados das diversas instituições selecionadas neste trabalho.

3.3 CARREGAMENTO E CONSULTA

Uma vez transformados em RDF/XML, foi necessário carregar os dados em um banco de dados específico para o armazenamento e a consulta de dados conectados, denominados, em inglês, como *triple-store*. Alguns exemplos incluem: Virtuoso, Apache Jena Fuseki, RDFFox e Neo4G. Em Rohloff *et al.* (2007), é apresentada uma análise de alguns desses sistemas. Neste trabalho, foi utilizado o Apache Jena Fuseki em razão de ele ter uma configuração mais simples e com suporte ao *Simple Protocol and RDF Query Language* (SPARQL). Esse protocolo permite que uma única consulta acesse várias bases de dados, tratando-as como se fossem um banco de dados global. Ele foi implantado na plataforma Heroku (www.heroku.com) e as consultas já podem ser realizadas no endereço http://sparql.dbacademic.tech/.

Na próxima seção, serão apresentados alguns resultados dessa primeira fase de desenvolvimento do repositório DBacademic, incluindo alguns exemplos de consultas.

4 – RESULTADOS

Foram extraídos sete conjuntos de dados de 25 instituições públicas de ensino, como detalhado na tabela 1. Observe que alguns conjuntos de dados estavam disponíveis em apenas algumas instituições.

Esses recursos foram então mapeados para conceitos, ou seja, classes de vocabulários já existentes. Os atributos também foram mapeados para propriedades e relacionados a termos de vocabulários, da mesma maneira, já existentes, como apresentados no quadro 4.

Como critério para a seleção dessas propriedades considerou-se a disponibilidade dessas informações nos diferentes portais. Mesmo assim, muitas delas não foram encontradas nos conjuntos de dados fornecidos pelas instituições, como o e-mail e o endereço para o currículo Lattes dos docentes.

Nessa fase, não foram desenvolvidos vocabulários específicos que fossem capazes de representar todos os dados das instituições de ensino. Focalizou-se apenas os dados extraídos e priorizou-se o reuso de vocabulários já existentes, entre eles, os apresentados no quadro 2 na seção 3.2. Contudo, alguns desses termos deverão ser revisados em trabalhos futuros, a exemplo do *Bibliographic Ontology Specification* (BIBO), uma ontologia bem consolidada para representar diversas publicações, incluindo dissertação de mestrado e tese de doutorado. Mesmo assim, nela não foi encontrado um conceito que pudesse representar com exatidão o que se denomina, no Brasil, trabalhos de conclusão de curso (ou monografias). Nessa fase de desenvolvimento, as monografias estão sendo mapeadas como bibo:Report, que não representa adequadamente o conceito de trabalho de conclusão de curso.

Na versão 0.1, a ontologia do DBacademic incluiu apenas os termos que foram reusados e o modo como eles se relacionam. Para melhor compreender as relações entre esses vocabulários, é possível consultar a versão mais atual no endereço http://purl.org/ontology/dbacademic.

Como exemplo, a figura 4 ilustra como uma monografia está relacionada a no mínimo outros quatro recursos.

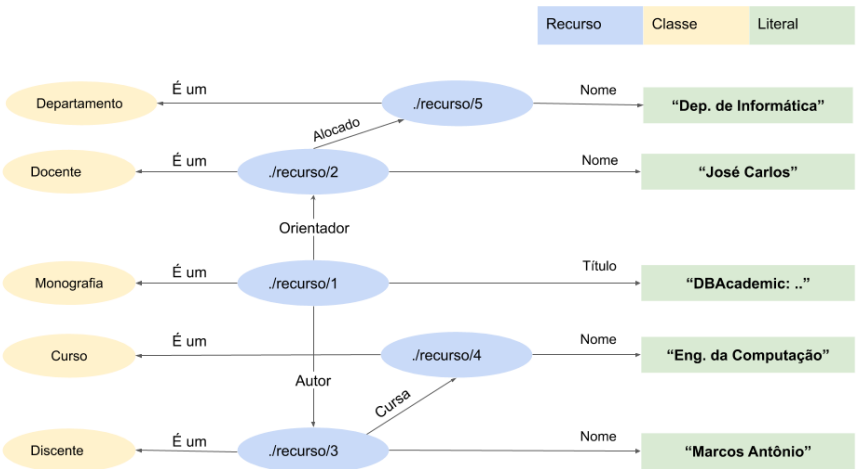
[4] Mais detalhes sobre essa biblioteca encontra-se em https://github.com/dbacademic/simpot

Tabela 1 – Recursos extraídos por instituição de ensino

Recursos	Instituições	Nº
Discente	UFMA, UFPI, UFRN, IFPA, IFMA, IFC, IFMS, IFRN, IFFAR	9
Docente	UFMA, UFPI, UFRN, UNIFESSPA, UFSJ, IFMA, IFPB, IFRN, IFMS, IFS, UFCSPA, UFV, UFMS, UFPEL, IFFAR	15
Curso	UFMA, UFPI, UFRN, UFPB, UFMS, IFMA, UFCA, UFCSPA, UFFS, UFPEL, IFMS, IFPB, IFRN, UFSJ, UFV, UNIFESSPA, UNIRIO, IFFAR	18
Centro	UFMA, UFPEL, UFRN,	3
Departamento	UFMA, UFRN, IFFAR	3
Grupo de pesquisa	UFRN, UFV, UFCA, IFC, UFPI, UFOP, UNIFESSPA, IFFAR, UFERSA	9
Monografia	UFMA, UFRN, UFOB	3
Total		60

Fonte: Elaboração do autor (2020).

Figura 4 – Representação gráfica da relação entre os recursos de uma monografia



Fonte: Elaboração do autor (2020).

Quadro 4 – Mapeamento de cada recursos para classes e propriedades de vocabulários existentes

Recurso	Classe	Propriedades
Docente	dbo:Professor	foaf:name, vcard:hasTelephone, vcard:hasPhoto, dbo:abstract, vcard:hasEmail, vcard:hasGender, owl:sameAs, cin:SIAPE, cin:academicDegree,cin:lattes, org:memberOf, owl:sameAs
Curso	aiiso:Programme	foaf:name, uai:hasKnowledgeArea, aiiso:responsibilityOf, aiiso:part_of, aiiso:code, owl:sameAs
Departamento	aiiso:Department	foaf:name, aiiso:responsibilityOf, owl:sameAs, aiiso:code, aiiso:part_of, owl:sameAs
Centro	aiiso:Center	foaf:name, aiiso:responsibilityOf,owl:sameAs, aiiso:code, owl:sameAs
Grupo de Pesquisa	aiiso:ResearchGroup	foaf:name, aiiso:hasKnowledgeArea, aiiso:responsibilityOf, owl:sameAs
Discente	cin:Student	foaf:name, dc:identifier, cin:isMemberOf, owl:sameAs
Monografia	bibo:Report	dc:title, dc:creator, bibo:issuer, dc:contributor, dc:date, owl:sameAs

Fonte: Elaboração do autor (2020).

Como ilustra a figura 4, o */recurso/1* é uma monografia que está relacionada a um docente *(./recurso/2)* e a um discente *(./recurso/3)*. Uma característica importante dos dados conectados é a possibilidade de um recurso ser facilmente conectado a recursos de outras bases de dados. O *./recurso/4,* por exemplo, poderia estar associado ao recurso http://pt.dbpedia.org/resource/Universidade_Federal_do_Maranhão, que pertence à base de dados do DBpedia.

Depois de extraídos e transformados, esses dados foram importados para um banco de dados conectado, resultando em 884.838 triplas. A tabela 2 apresenta a quantidade de triplas por cada classe.

Tabela 2 – Quantidade de triplas associadas à cada classe

Classe	Nº Triplas
cin:Student	132.936
bibo:Report	61.875
dbo:Professor	24.564
aiiso:ResearchGroup	2.917
aiiso:Programme	2.367
aiiso:Department	956
aiiso:Center	53

Fonte: Elaboração do autor (2020).

As instituições que tiveram mais dados carregados foram: UFRN, IFRN e UFMA. A tabela 3 apresenta as dez instituições que possuem mais dados.

Tabela 3 – Número de triplas por instituição

Instituição	Nº Triplas
UFRN	71415
IFRN	41173
UFMA	40133
IFMA	23565
UFPA	15950
IFFAR	10019
UFV	4178
UFPI	3751
UFSJ	3445
UFMS	2234

Fonte: Elaboração do autor (2020).

Os dados conectados são consultados por meio de um SPARQL *endpoint*, que é provido pelo servidor de dados conectados, ou *triple store*. Para este trabalho, os dados foram carregados e implantados no servidor Apache Jena Fuseki e estão disponíveis no endereço http://sparql.dbacademic.tech/.

Através deste endereço, é possível escrever e enviar as consultas diretamente, usando um navegador Web. Como exemplo, o quadro 5 apresenta um código que irá retornar ordenadamente a quantidade de cursos de engenharia por instituição de ensino. Essa consulta pode ser acessada também no endereço encurtado https://bit.ly/35lAA7d.

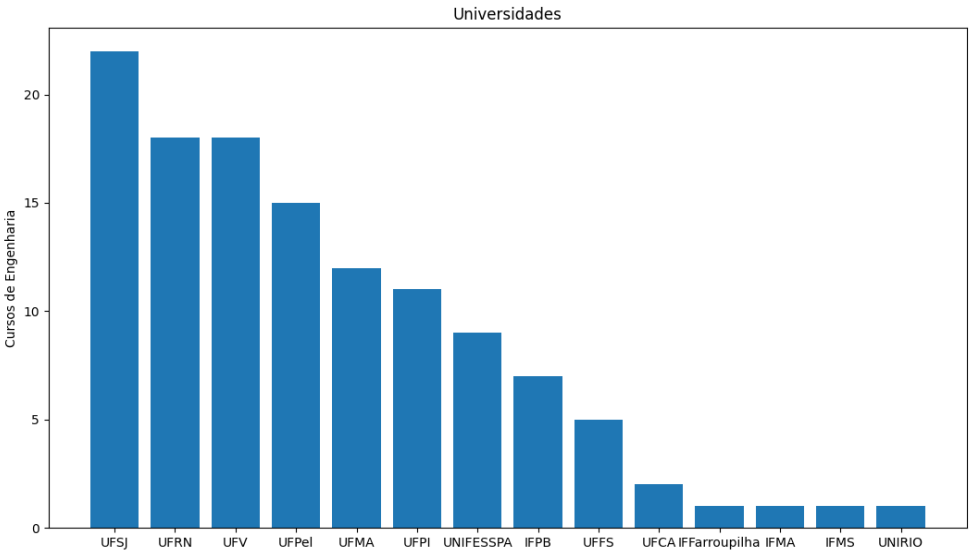
Quadro 5 – Consulta SPARQL que retorna a quantidade de cursos de engenharia por instituição

```
PREFIX dbo: <http://dbpedia.org/ontology/>

PREFIX aiiso: <http://purl.org/vocab/aiiso/schema#>

PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT ?sigla (COUNT(?curso) as ?quant_cursos)
WHERE {

?curso a aiiso:Programme.

?curso aiiso:part_of ?instituicao.

?instituicao foaf:nick ?sigla.

?instituicao a dbo:EducationalInstitution.

FILTER regex(lcase(?curso_name), "engenharia") }

GROUP BY ?sigla

ORDER BY DESC (?quant_cursos)
```

Fonte: Elaboração do autor (2020).

Além da possibilidade de consultar os dados diretamente pelo navegador Web, é possível enviar as consultas por intermédio de um aplicativo móvel ou de um algoritmo de análise e visualização de dados. A título de exemplo, um código escrito na linguagem Python pode enviar a consulta do quadro 5 e gerar um gráfico a partir do resultado. Esse código pode ser acessado e executado diretamente pelo Google Colab (https://colab.research.google.com/) no endereço https://bit.ly/2y2aTfZ. Ao executar esse código é gerada a saída apresentada na figura 5.

Figura 5 – Gráfico produzido de uma consulta SPARQL ao DBAcademic



Fonte: Elaboração do autor (2020).

A figura 5 apresenta um gráfico de barras com a quantidade de cursos de engenharia por instituição de ensino, apontando a UFSJ com a maior quantidade de cursos de Engenharia, lembrando que esse resultado é gerado a partir dos dados que estão no DBacademic e que não inclui ainda todas as instituições de ensino.

Uma das vantagens dos dados conectados, discutidos na seção 2, é a possibilidade de incluir os dados de outras bases de dados, como o DBpedia. No exemplo do quadro 6, uma instituição na base de dado do DBacademic é associada a um recurso na DBpedia. Com essa associação, é possível chegar a propriedade dbo:city, que levu a outro recurso, que permitu acessar o nome dessa cidade. Esse nome é então inclnído nos resultados dessa consulta, que é realizada a partir do SPARQL *endpoint* do DBacademic.

Quadro 6 – Consulta SPARQL que permite a integração de bases de dados conectados

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
SELECT distinct ?instituicao ?nome_cidade
WHERE {
  ?instituicao a dbo:EducationalInstitution.
  ?instituicao owl:sameAs ?same.
  SERVICE <http://dbpedia.org/sparql> {
    ?same dbo:city ?cidade.
    ?cidade foaf:name ?nome_cidade    }
  FILTER REGEX(str(?same), "^http://dbpedia.org")}
```

Fonte: Elaboração do autor (2020).

Do mesmo modo que, no quadro 6, foi incluído o DBpedia, seria possível incluir diversas outras bases de dados, enriquecendo as análises e a visualização dos dados.

Por meio desses resultados, é possível identificar o potencial do DBacademic de integração de dados, valorizando ainda mais os dados que já estão abertos pelas diversas instituições de ensino. Contudo, para a sua ampliação, deverão ser considerados os três grandes desafios a seguir.

A NECESSIDADE DE ELABORAÇÃO DE UMA ONTOLOGIA MAIS ADEQUADA

A criação de ontologias é sempre um grande desafio, e poderá ser muito beneficiada com a participação de outras instituições de ensino. No Brasil, foram identificados dois projetos que avançaram na proposta de uma ontologia que faz o reuso de diversos vocabulários (ALENCAR *et al.*, 2018; ROCHA; LÓSCIO, 2015). Desse modo, para o próximo passo, será necessário compatibilizar e adequar melhor as propostas existentes para atender às demandas do Dbacademic.

A MELHORIA NA QUANTIDADE E QUALIDADE DOS DADOS

Mesmo sem realizar uma análise quantitativa e qualitativa dos portais de dados abertos, foi possível perceber a ausência e a falta de atualização em muitos deles. Em geral, o que está disponível nos portais são apenas visões dos dados, definidas pelo detentor deles. No paradigma de dados conectados, armazenam-se apenas os dados e suas conexões. As visões são criadas por meio de consultas, que poderiam incluir dados de outras bases, como destacado na seção 4.

A NECESSIDADE DE UM PLANO PARA A MANUTENÇÃO E INSTITUCIONALIZAÇÃO DO PROJETO

Na literatura, existem diversas iniciativas similares e relevantes que atualmente estão tendo dificuldades de serem mantidas. Muitos portais de dados conectados já não estão mais em funcionamento e suas ontologias não estão mais disponíveis. Essa dificuldade pode ser ainda maior, devido à necessidade de manter atualizados os dados de diversas instituições.

5 – CONCLUSÕES

Este artigo apresentou um projeto que tem o objetivo de conectar dados abertos de diversas instituições em um grande repositório de dados, denominado DBacademic. Os resultados dessa primeira fase do projeto mostraram o grande potencial dessa proposta, que poderá se tornar uma importante referência, principalmente, para consultas a dados entre instituições de ensino.

Como resultados dessa primeira fase, o DBacademic já conseguiu incluir sete conjuntos de dados de 25 instituições de ensino do Brasil. Atualmente, já é possível realizar diversas consultas em sua base de dados com quase 900 mil triplas. Não foi realizada ainda qualquer avaliação de eficiência dessas consultas, visto que o servidor atualmente implantado utiliza uma quota gratuita com algumas limitações de eficiência. Espera-se que, ao finalizar essa fase de teste, consigamos implantar o repositório em um servidor dedicado e mantido por uma das instituições de ensino.

Além dos resultados, deve-se destacar os três principais desafios: necessidade de elaboração de uma ontologia mais adequada; melhoria na qualidade dos dados; e necessidade de um plano para a manutenção e institucionalização do projeto. Todos esses desafios poderão ser enfrentados por meio de parcerias com instituições e pesquisadores. Nesse sentido, um dos objetivos subjacentes deste artigo é apresentar e validar esse projeto com a comunidade acadêmica para fazer avançar as parcerias.

Além dos três desafios, um estudo relevante a ser considerado em um trabalho futuro é a análise da efetividade dos Planos de Dados Abertos elaborados pelas instituições, incluindo a identificação de quais desafios elas estão enfrentando para seguir os oito princípios fundamentais dos dados abertos propostos pelo open Government Working Group (2007).

REFERÊNCIAS

ALENCAR, A.; XAVIER, D.; CHAVES, L. C.; SOUZA, D. Y. Publicação e consumo de dados abertos conectados acadêmicos. *Revista Principia:* Divulgação Científica e Tecnológica do IFPB, v. 1, n. 42, p. 136, 18 ago. 2018. DOI 10.18265/1517-03062015v1n42p136-145. Disponível em: http://periodicos.ifpb.edu.br/index.php/principia/article/view/1988. Acesso em: 11 mar. 2021.

AUER, S.; BIZER, C.; KOBILAROV, G.; LEHMANN, J.; CYGANIAK, R.; IVES, Z. DBpedia: A Nucleus for a Web of Open Data. *In*: ABERER, K.; CHOI, K.-S.; NOY, N.; ALLEMANG, D.; LEE, K.-I.; NIXON, L.; GOLBECK, J.; MIKA, P.; MAYNARD, D.; MIZOGUCHI, R.; SCHREIBER, G.; CUDRÉ-MAUROUX, P. (orgs.). *The Semantic Web*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. v. 4825, p. 722–735. DOI 10.1007/978-3-540-76298-0_52. Disponível em: http://link.springer.com/10.1007/978-3-540-76298-0_52. Acesso em: 11 mar. 2021.

BERNERS-LEE, T. *Linked Data*. 2009. Disponível em: https://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 13 jan. 2021.

BERTIN, P. R. B.; MACHADO, C. D.; VISOLI, M. C.; DRUCKER, D. P.; PINTO, D. M. A construção do Plano de Dados Abertos de uma organização pública de Pesquisa e Desenvolvimento e o desafio de uma Ciência Agropecuária Aberta. *Revista Eletrônica de Comunicação, Informação e Inovação em Saúde*, v. 11, 30 nov. 2017. DOI 10.29397/reciis.v11i0.1411. Disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/1411. Acesso em: 11 mar. 2021.

BRASIL. *Decreto nº 8.777, de 11 de maio de 2016.* Institui a Política de Dados Abertos do Poder Executivo federal. 11 maio 2016. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm. Acesso em: 22 fev. 2021.

BRASIL. *Lei nº 12.527, de 18 de novembro de 2011 [Lei de Acesso à Informação]*. Regula o acesso a informações previsto no inciso XXXIII do art. 5º , no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. 2011. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 30 mar. 2020.

BRASIL. *Portal Brasileiro de Dados Abertos*. 2019. Disponível em: http://dados.gov.br. Acesso em: 13 set. 2019.

BRASIL. MINISTÉRIO DA DEFESA. Sobre a Lei de Acesso à Informação. 9 set. 2020. *Gov.br*. Disponível em: https://www.gov.br/defesa/pt-br/acesso-a-informacao/outros/sobre-lei-de-acesso-a-informacao. Acesso em: 11 mar. 2021.

BREITMAN, K. K. *Web semântica: a internet do futuro*. Rio de Janeiro: LTC, 2006.

CAROSSI, D. F.; TEIXEIRA FILHO, J. G. de A. Uma Análise dos Pedidos de Acesso à Informação Encaminhados a uma Instituição de Ensino Superior. *Gestão.Org*, v. 14, n. 2special, p. 255–264, 1 maio 2017. DOI 10.21714/1679-18272016v14Esp2.p255-264. Disponível em: http://www.revista.ufpe.br/gestaoorg/index.php/gestao/article/viewFile/906/528. Acesso em: 11 mar. 2021.

COSTA, I. N. da; ANDRADE, L. do E. S.; RESENDE, L.; TONIN, P.; COSTA, M.; SANTOS, Z. *Manual da Lei de Acesso à Informação para Estados e Municípios*. Brasília: CGU, 2013(Brasil transparente). Disponível em: https://acessoainformacao.valparaisodegoias.go.gov.br/res/docs/manual_lai_estadosmunicipios.pdf. Acesso em: 11 mar. 2021.

COSTA, S. S.; SOUSA, M. V. D.; SILVA, M. L. da; OLIVEIRA, E. C. de; GUIMARÃES, J. V. M. Uma solução semi-automática para extração, transformação e carga de dados abertos conectados. *In*: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 2019. *Anais* [...]. Brasília: FCI, 2019. p. 138–143. Disponível em: http://widat2019.fci.unb.br/index.php/anais-widat-2019. Acesso em: 11 mar. 2021.

GAMA, J. R.; RODRIGUES, G. M. Transparência e acesso à informação: um estudo da demanda por informações contábeis nas universidades federais brasileiras. *Transinformação*, v. 28, n. 1, p. 47–58, abr. 2016. DOI 10.1590/2318-08892016002800004. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-37862016000100047&lng=pt&tlng=pt. Acesso em: 11 mar. 2021.

HEATH, T.; BIZER, C. 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, 9 fev. 2011. DOI 10.2200/S00334ED1V01Y201102WBE001. Disponível em: http://www.morganclaypool.com/doi/abs/10.2200/S00334ED1V01Y201102WBE001. Acesso em: 11 mar. 2021.

ISOTANI, S.; BITTENCOURT, I. I. *Dados abertos conectados*. São Paulo: Novatec, 2015.

KESSLER, C.; KAUPPINEN, T. Linked Open Data University of Münster – Infrastructure and Applications. *In*: SIMPERL, E.; NORTON, B.; MLADENIC, D.; DELLA VALLE, E.; FUNDULAKI, I.; PASSANT, A.; TRONCY, R. (orgs.). *The Semantic Web: ESWC 2012 Satellite Events*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015. v. 7540, p. 447–451. DOI 10.1007/978-3-662-46641-4_43. Disponível em: http://link.springer.com/10.1007/978-3-662-46641-4_43. Acesso em: 11 mar. 2021.

MCCRAE, J. P.; ABELE, A.; BUITELAAR, P.; CYGANIAK, R.; JENTZSCH, A.; ANDRYUSHECHKIN, V.; DEBATTISTA, J.; NASIR, J. *The Linked Open Data Cloud.* 20 maio 2020. Disponível em: https://lod-cloud.net/. Acesso em: 11 mar. 2021.

OPEN GOVERNMENT WORKING GROUP. *The 8 Principles of Open Government Data*. 2007. Disponível em: https://public.resource.org/8_principles.html. Acesso em: 11 mar. 2021.

OPEN KNOWLEDGE FOUNDATION. *Definição de Conhecimento Aberto*. 2019. Disponível em: https://opendefinition.org/od/2.0/pt-br/. Acesso em: 9 set. 2020.

PIEDRA, N.; TOVAR, E.; COLOMO-PALACIOS, R.; LOPEZ-VARGAS, J.; ALEXANDRA CHICAIZA, J. Consuming and producing linked open data: the case of OpenCourseWare. *Program*, v. 48, n. 1, p. 16–40, 28 jan. 2014. DOI 10.1108/PROG-07-2012-0045. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/PROG-07-2012-0045/full/html. Acesso em: 11 mar. 2021.

RIVEST, R. *The MD5 Message-Digest Algorithm*, n. RFC1321. [*S. l.*]: RFC Editor, abr. 1992. DOI 10.17487/rfc1321. Disponível em: https://www.rfc-editor.org/info/rfc1321. Acesso em: 11 mar. 2021.

ROCHA, J.; LÓSCIO, B. OpenCIn: Usando Dados Abertos e Conectados para a Publicação de dados sobre o CIn/UFPE. *In*: CONCURSO DE TRABALHOS DE INICIAÇÃO CIENTÍFICA DA SBC (CTIC-SBC), 34., 2015. *Anais* [...]. Recife: Sociedade Brasileira de Computação, 2015. v. 34, p. 11–20. Disponível em: https://sol.sbc.org.br/index.php/ctic/article/view/10014. Acesso em: 11 mar. 2021.

ROHLOFF, K.; DEAN, M.; EMMONS, I.; RYDER, D.; SUMNER, J. An Evaluation of Triple-Store Technologies for Large Data Stores. *In*: MEERSMAN, R.; TARI, Z.; HERRERO, P. (org.). *On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshops*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. v. 4806, p. 1105–1114. DOI 10.1007/978-3-540-76890-6_38. Disponível em: http://link.springer.com/10.1007/978-3-540-76890-6_38. Acesso em: 11 mar. 2021.

TORINO, E.; TREVISAN, G. L.; VIDOTTI, S. A. B. G. Os diferentes conceitos de dados de pesquisa na abordagem da biblioteconomia de dados. *Ciência da Informação*, v. 48, n. 3 (Supl.), p. 38–46, dez. 2019. Disponível em: http://revista.ibict.br/ciinf/article/view/4866/4428. Acesso em: 11 mar. 2021.

VASSILIADIS, P.; SIMITSIS, A.; SKIADOPOULOS, S. Conceptual modeling for ETL processes. *In*: THE 5TH ACM INTERNATIONAL WORKSHOP, 2002. *Proceedings* [...]. McLean, Virginia, USA: ACM Press, 2002. p. 14–21. DOI 10.1145/583890.583893. Disponível em: http://portal.acm.org/citation.cfm?doid=583890.583893. Acesso em: 11 mar. 2021.

ZABLITH, F.; FERNANDEZ, M.; ROWE, M. The OU Linked Open Data: Production and Consumption. *In*: GARCÍA-CASTRO, R.; FENSEL, D.; ANTONIOU, G. (orgs.). *The Semantic Web: ESWC 2011 Workshops*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. v. 7117, p. 35–49. DOI 10.1007/978-3-642-25953-1_4. Disponível em: http://link.springer.com/10.1007/978-3-642-25953-1_4. Acesso em: 11 mar. 2021.

ZORZAL, L.; RODRIGUES, G. M. Transparência das informações das universidades federais: estudo dos relatórios de gestão à luz dos princípios de governança. *Biblios:* Journal of Librarianship and Information Science, n. 61, p. 1–18, 14 mar. 2016. DOI 10.5195/BIBLIOS.2015.253. Disponível em: http://biblios.pitt.edu/ojs/index.php/biblios/article/view/253. Acesso em: 11 mar. 2021.

AGRADECIMENTOS

Agradecemos aos autores dos projetos OpenCIN e OpenUAI pela comunicação, que foi breve, mas essencial a este trabalho. Espera-se que este trabalho seja apenas um ponto inicial para futuras parcerias

Acervos Culturais Brasileiros no Repositório Wikimedia Commons: um estudo sobre o reúso e a visualização de mídias referentes a coleções de museus do Instituto Brasileiro de Museus(Ibram)

Danielle do Carmo
Doutoranda em Ciência da Informação pela Universidade de Brasília (UnB) - Brasília, DF - Brasil. Mestre em Memória Social e Patrimônio Cultural pela Universidade Federal de Pelotas (UFPEL) – RS - Brasil.
http://lattes.cnpq.br/6139212172511823
E-mail:docarmo.danielle@gmail.com

Dalton Lopes Martins
Pós-Doutorado pela Universidade de São Paulo (USP) – SP - Brasil. Doutor em Ciência da Informação pela Universidade de São Paulo (USP) - São Paulo, SP - Brasil. Professor da Universidade de Brasília (UnB) - Brasília, DF - Brasil.
http://lattes.cnpq.br/3774617443225038
E-mail:dmartins@gmail.com

Submetido em: 23/09/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021.

RESUMO

O artigo apresenta um estudo webométrico acerca de mídias referentes a coleções de acervos de museus brasileiros no repositório de mídias Wikimedia Commons, mais especificamente sobre mídias de coleções de nove museus geridos pelo Instituto Brasileiro de Museus (Ibram). Utilizando as ferramentas GLAMorous e GLAMorous 2, foi possível obter dados relativos ao reúso de mídias da Wikimedia Commons em outras plataformas wiki da Fundação Wikimedia, como os projetos Wikipédia de diferentes idiomas, a Wikidata e outros. Também foi possível obter números relativos à quantidade de visualizações desses acervos no ano de 2019. Dessa forma a investigação realizada, assim como os dados obtidos, podem ajudar instituições culturais guardiãs de acervos, como os museus, a entender o que acontece com seus acervos digitais uma vez que são disponibilizados nas plataformas wiki, fornecendo dados importantes para a gestão de suas coleções, principalmente em relação ao acesso e reapropriação desses acervos por parte dos usuários.

Palavras-chave: Wikimedia Commons. Acervos culturais digitais. Wikidata. Instituições culturais. Museus. GLAM.

Brazilian Cultural Collections in the Wikimedia Commons Repository: a study on the reuse and visualization of media related to museum collections of the Brazilian Institute of M

useums (Ibran)

ABSTRACT

This article presents a webometric study about media referring to collections of Brazilian museums in the Wikimedia Commons media repository, more specifically about media from collections of nine museums managed by the Brazilian Museum Institute (Ibram). Using the GLAMorous and GLAMorous 2 tools, it was possible to obtain data regarding the media reuse of Wikimedia Commons in other Wikimedia Foundation wiki platforms, such as Wikipedia projects in different languages, Wikidata and others. It was also possible to obtain numbers related to the views of these collections in 2019. In this way, the research carried out, as well as the data obtained, ways to help cultural institutions guarding collections, such as museums, to understand what happens with their digital collections. Since they are made available on wiki platforms, providing important data for the management of their collections, mainly in relation to access and reappropriation of these collections by users.

Keywords*: Wikimedia Commons. Digital cultural collections. Wikidata. Museums. Cultural institutions. GLAM.*

Colecciones culturales brasileñas en el repositorio de Wikimedia Commons: un estudio sobre la reutilización y visualización de medios relacionados con colecciones de museos del Instituto Brasileño de Museos(Ibran)

RESUMEN

Este artículo presenta un estudio webométrico sobre los medios que se refieren a colecciones de museos brasileños en el repositorio de medios de Wikimedia Commons, más específicamente sobre medios de colecciones de nueve museos administrados por el Instituto Brasileño de Museos (Ibram). Utilizando las herramientas GLAMorous y GLAMorous 2, fue posible obtener datos sobre la reutilización de medios de Wikimedia Commons en otras plataformas de wikis de la Fundación Wikimedia, como proyectos de Wikipedia en diferentes idiomas, Wikidata y otros. También fue posible obtener números relacionados con el número de vistas de estas colecciones en 2019. De esta manera, la investigación llevada a cabo, así como los datos obtenidos, pueden ayudar a las instituciones culturales que custodian colecciones, como los museos, a comprender lo que sucede con sus colecciones digitales. ya que están disponibles en plataformas wiki, proporcionando datos importantes para la gestión de sus colecciones, principalmente en relación con el acceso y la reapropiación de estas colecciones por parte de los usuarios.

Palabras clave*: Wikimedia Commons. Colecciones culturales digitales. Wikidata. Instituciones culturales. Museos. GLAM.*

INTRODUÇÃO

Com o surgimento das novas tecnologias de informação e comunicação, mais especificamente com o advento da internet, pôde-se observar a emergência de novos espaços sociais baseados no digital e em novos tipos de fluxos informacionais. Esses espaços, ocupados com objetivos e propósitos diversos, oferecem aos seus usuários meios de consumo e compartilhamento de informações, mecanismos de interação com as informações e dispositivos que promovem a conexão e comunicação direta entre usuários. Como exemplo desses meios sociais digitais que emergiram principalmente com o surgimento da Web 2.0, podemos citar as plataformas baseadas em conteúdos gerados pelo usuário, como a rede social Facebook, a rede de micro blogs Twitter, o streaming de vídeo YouTube e a enciclopédia livre Wikipédia. Esses espaços sociotécnicos, além de serem entendidos como meios potenciais de socialização da informação, também podem ser entendidos como fontes de informação que permitem a coleta de dados sobre a circulação e a apropriação de objetos digitais, que nos fornece informações sobre como são disponibilizados, contextualizados e descritos.

Nesse sentido, a webometria pode ser uma importante aliada quando utilizada em estudos que pretendem investigar fenômenos baseados na web. Os autores Björneborn e Ingwersen (2004) definiram a webometria como o "estudo de fenômenos da web baseados em técnicas quantitativas e recorrendo a métodos infométricos" (2004). Para Thelwall (2009, p. 1) essa definição foi importante por atribuir à webometria características de método infométrico, pois dessa forma os autores a posicionaram como um campo da ciência da informação. Entretanto, o autor propõe uma definição mais abrangente ao dizer que a webometria é "o estudo dos conteúdos da web com métodos essencialmente quantitativos que não são específicos para um campo de estudo" (THELWALL, 2009, p. 6). Vanti (2002, p. 161) posiciona a webometria, a bibliometria e a cienciometria como subcampos da infometria e diz que todas "têm funções semelhantes, mas ao mesmo tempo cada uma delas propõe medir a difusão do conhecimento científico e o fluxo da informação sob enfoques diversos" (VANTI, 2002). Entre os estudos webométricos estariam incluídos aqueles que adotam métodos de rastreamento das ações dos usuários on-line que, por meio de ferramentas de análise, podem captar informações que nos permitem medir aspectos do uso de recursos disponíveis na web (THELWALL, 2009, p. 89).

Do mesmo modo que os estudos bibliométricos permitem melhor administração das coleções no contexto das bibliotecas, os estudos webométricos podem dar subsídios que auxiliem as instituições culturais como os museus a administrar seus recursos, entre eles objetos digitais de acervos culturais, e entender como se dá seu uso pelos usuários que a eles têm acesso. Logo, a presente investigação explora algumas ferramentas de coleta de dados e apresenta possibilidades de análise e mensuração de aspectos relacionados ao acesso e à reutilização de acervos brasileiros que estão disponíveis na web por meio do repositório de mídias Wikimedia Commons.

O repositório de mídias on-line Wikimedia Commons armazena e disponibiliza, gratuitamente, diversos tipos de arquivos de mídias que são fornecidos de forma coletiva e colaborativa por meio da ação de usuários voluntários. É um projeto da organização sem fins lucrativos Fundação Wikimedia, que também é a responsável pela manutenção de outros projetos baseados em conteúdo gerado por usuários como a Wikipédia, o Wikitionary, o Wikibooks, o Wikisource, o Wikinews, o Wikiversite, o Wikiquote, o Wikidata e outros.

O Wikimedia Commons, lançado no início de setembro de 2004, disponibiliza atualmente mais de 60 milhões de arquivos de mídias. Essas mídias são disponibilizadas sob licenças individuais que permitem a cópia, a reutilização e a modificação, de acordo com termos especificados.

Os conteúdos disponibilizados por meio do repositório Wikimedia Commons estão sob a licença Creative Commons Attribution / Share-Alike (WIKIMEDIA COMMONS, 2020a). Segundo dados fornecidos pelo projeto, em abril de 2020, a plataforma registra um total de 43.020 usuários ativos e 168 *bots, programas que podem editar ou submeter conteúdo para a plataforma de modo automático,* que auxiliam na edição do conteúdo (WIKIMEDIA COMMONS, 2020b).

Ao explorar o repositório Wikimedia Commons, é possível observar que estão disponíveis diferentes tipos mídias de diversas temáticas. Entre as mídias é possível identificar conteúdos referentes a itens de acervos oriundos de coleções de diversas instituições culturais como museus, arquivos e galerias localizadas em diversas partes do mundo. Essas mídias podem ter sido disponibilizadas tanto por usuários que realizaram o upload desse material de modo espontâneo, ou em alguns casos por meio de parcerias estabelecidas entre usuários e instituições culturais. Segundo os autores Stinson, Fauconnier e Wyatt (2018, p. 17), há um histórico de colaboração entre a comunidade Wikimedia e as instituições culturais, reunidas sob o termo guarda-chuva GLAM[1]. Essas instituições identificaram nos projetos da Fundação Wikimedia, em especial no Wikimedia Commons, na enciclopédia on-line Wikipédia e no repositório de dados estruturados Wikidata, meios potenciais de difusão de seus acervos na internet. Para se ter uma noção dos efeitos dessa prática, Zeinstra (2013) aponta que, no ano 2013, o conteúdo GLAM reunia cerca de mais de dois milhões de objetos digitais no Wikimedia Commons, o que corresponderia ao total de 13,14% do conteúdo da plataforma.

> Em outras palavras, um em cada oito arquivos na Wikimedia Commons é disponibilizado através de alguma colaboração com GLAM's. Tanto a GLAM coloca suas coleções em domínio público ou abre a licença de suas coleções on-line e voluntários realizam upload, ou alguma colaboração ativa entre GLAM's e a comunidade Wikimedia como Wiki Loves Monuments cria conteúdo GLAM. Dessa forma, voluntários colocam esses objetos de mídia em artigos da Wikipédia criando um incrível aumento em sua visibilidade. Instituições contribuintes enxergam o potencial da Wikimedia como um canal de distribuição (ZEINSTRA, 2013, p. 5, tradução nossa).

As autoras Villaespesa e Navarete (2019) chamam a atenção para a utilização de mecanismos de busca como o Google e o Google Imagens, ou da busca de voz de assistentes virtuais como a Siri e a Alexa é possível identificar uma exposição privilegiada de conteúdos da Wikimedia Commons, Wikidata e Wikipédia. Essa exposição privilegiada na ordem de apresentação de resultados se daria devido "aos ambientes estruturados fornecidos pelas plataformas Wiki que influenciam fortemente os resultados das pesquisas" (VILLAESPESA; NAVARRETE, 2019).

Além de se apresentar como potenciais aliados na difusão das informações referentes aos acervos das instituições culturais, os projetos da Fundação Wikimedia constituem-se como espaços sociotécnicos que promovem práticas de curadoria coletiva e que podem nos fornecer dados relevantes sobre aspectos relacionados ao acesso e à reutilização das coleções de acervos culturais compartilhados nesses ambientes. Nesse sentido a Fundação Wikimedia, por meio de uma página na Wikimedia Outreach sobre os projetos com GLAM's, indica uma série de ferramentas desenvolvidas com o objetivo de captar e os expor dados sobre as mídias digitais de coleções culturais em ambientes como a Wikimedia Commons (WIKIMEDIA OUTREACH, 2020).

O objetivo da presente pesquisa foi explorar como esses dados podem ser extraídos e analisados, evidenciando seu potencial como fonte de informação sistematizada, assim como ampliar a própria sensibilização da área da ciência da informação para os fenômenos culturais que se dão nesses ambientes e o quanto eles podem indicar novas práticas sociais de gestão e curadoria da informação que devem ser levadas em consideração em pesquisas na área.

[1] Acrônimo em inglês - Galleries, Library, Archives and Museums - usado para se referir a instituições como galerias, bibliotecas, arquivos, museus e outras instituições do patrimônio cultural.

O foco da pesquisa é investigar a presença de acervos culturais na Wikimedia Commons e entender como se configura essa presença a partir de diferentes tipos de indicadores e informações disponibilizadas pela plataforma. A amostra utilizada foi de mídias de coleções de museus que estão sob a gerência do Instituto Brasileiro de Museus (Ibram).

METODOLOGIA

A pesquisa se apresenta como exploratória e descritiva, de natureza quantitativa, e para que fosse possível coletar os dados para análise, utilizaram-se duas ferramentas, a GLAMorous (MANSKE, 2020a) e a GLAMorous 2 (MANSKE, 2020b). Ambas as ferramentas foram desenvolvidas pelo wikimedista2 Magnus Manske e se encontram listadas em uma página de ferramentas voltadas para os GLAM's no Wikimedia Outreach (WIKIMEDIA OUTREACH, 2020). Com essas ferramentas foi possível obter dados de visualização dos arquivos das mídias do Wikimedia Commons com base nas categorias. Assim, depois de identificar os museus sob gestão do Ibram (2020), exploramos as categorias referentes às coleções destes museus na Wikimedia Commons. Uma vez identificadas as categorias, foi possível utilizá-las nos mecanismos de busca das ferramentas selecionadas e estabelecer os parâmetros. A ferramenta GLAMorous permite o rastreamento de uso de imagens que estão em uma categoria do Wikimedia Commons, e por meio dela foi possível obter dados relativos ao reúso dessas mídias em páginas de diversos projetos da Fundação Wikimedia. É importante destacar que a possibilidade de rastrear o uso de objetos digitais pode ser aplicada em pesquisas que tenham por objetivo compreender as formas de apropriação social e reúso da informação, sobretudo ressaltando o contexto em que elas são utilizadas; desse modo se torna possível analisar e evidenciar significados fundamentais para a compreensão dos fluxos informacionais constituídos nesses ambientes. Desde o momento de publicação de um objeto digital por uma instituição, passando pelo momento em que esse objeto é escolhido para ilustrar um conceito em um verbete, até sua visualização por um usuário em um verbete específico na Wikipédia, diversas camadas podem compreendidas por meio da obtenção dessas informações registradas, podendo assim apoiar a compreensão do fenômeno social de uso da informação que ainda hoje carece de entendimento e fortalecimento metodológico. A ferramenta GLAMorous 2 complementa a GLAMorous, apresentando funcionamento semelhante, porém, além de dados de reúso, é possível obter dados de visualização dos arquivos. Com essas ferramentas foi possível realizar as análises webométricas que serão apresentados nas próximas seções do presente artigo.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

A REUTILIZAÇÃO DE IMAGENS DE COLEÇÕES CULTURAIS NA WIKIMEDIA COMMONS EM OUTROS PROJETOS DA FUNDAÇÃO WIKIMEDIA

Ao identificar os 30 museus que estão sob a administração direta do Ibram, foi possível localizar categorias de mídias que correspondem às suas coleções no Wikimedia Commons por meio de buscas na plataforma[3]. Assim identificamos as categorias referentes às coleções de nove museus que estão elencadas no quadro 1.

Identificadas as categorias referentes a coleções de museus, foi possível utilizá-las na realização de buscas na ferramenta GLAMorous, e assim coletar estatísticas atuais de reúso de imagens da Wikimedia Commons em páginas da Wikipédia em português, em Wikipédias de outros idiomas e em páginas de outros projetos da Fundação Wikimedia, como o Wikibooks, Wikiquotes e Wikidata. Além disso foi possível obter a quantidade de imagens encontradas em cada categoria, o total de imagens distintas utilizadas, o total de uso das imagens das categorias, assim como a porcentagem do total de imagens da categoria que estão sendo reutilizadas, como pode ser visto nas tabelas 1 e 2.

[2] Nome atribuído aos usuários e colaboradores dos diversos projetos da Fundação Wikimedia.

[3] Busca por categorias: Disponível em: <https://commons.wikimedia.org/wiki/Special:Categories>. Acesso em abr. 2020.

Quadro 1 – Categorias referentes às coleções dos museus no Wikimedia Commons

#	Museu	Categoria
1	Museu da Inconfidência	Collections of the Museu da Inconfidência
2	Museu da República	Collections of the Museu da República
3	Museu Histórico Nacional	Collections of the Museu Histórico Nacional
4	Museu Imperial	Collections of the Museu Imperial
5	Museu Nacional de Belas Artes	Collections of the Museu Nacional de Belas Artes
6	Museu do Açude (equipamento dos Museus Castro Maya)	Collections of the Museu Castro Maya
7	Museu Casa de Benjamin Constant	Media contributed by the Museu Casa de Benjamin Constant
8	Museu Regional de São João del-Rei	Collections of the Museu Regional de São João del-Rei
9	Museu Victor Meirelles	Collections of the Museu Victor Meirelles

Fonte: Dados da pesquisa, 2020.

Com base nos dados obtidos, é possível observar que a coleção de museu que mais apresenta imagens reutilizadas na ilustração páginas da Wikipédia em português é a categoria referente ao Museu da República, que reúne 33 ocorrências de imagens presentes em artigos.

Essa categoria também é a que ilustra mais Wikipédias de outros idiomas, mostrando assim 75 ocorrências em páginas de Wikipédias em outros idiomas, uma ocorrência de imagem no Wikibooks e uma no Wikiquote. Assim foi possível constatar que das nove imagens encontradas na categoria correspondente à coleção do Museu da República, 4 estão sendo reutilizadas em 110 páginas de projetos da Fundação Wikimedia, o que corresponde ao total de uso de 44,4% das imagens disponíveis na categoria.

A categoria referente à coleção do Museu Histórico Nacional revelou o maior número de arquivos, com 471 imagens encontradas na categoria. Essas imagens foram reutilizadas 13 vezes em páginas da Wikipédia em português e 19 vezes em 7 páginas de Wikipédias de outros idiomas. Das categorias selecionadas para a pesquisa foi a que apresentou maior ocorrência de imagens, ilustrando itens no Wikidata com 4 ocorrências. Das 471 imagens disponibilizadas na categoria, há 9 imagens sendo reutilizadas em 36 ocorrências, o que constitui 2,55% de reúso do total de imagens disponíveis na categoria.

A coleção referente ao Museu Castro Maya é a categoria que apresenta maior quantidade de imagens distintas utilizadas. A categoria disponibiliza o total de 46 imagens, das quais 17 imagens indicaram ocorrência em 12 páginas da Wikipédia em português, em 25 páginas de Wikipédias de outros idiomas, em cinco páginas do Wikibooks e em três itens do Wikidata, somando assim 45 ocorrências de utilização de imagens, o que corresponde ao um reúso de 36,96% de todas as imagens disponíveis na categoria.

Apesar de apontar somente sete imagens na categoria referente ao Museu da Inconfidência, essa coleção é a que apresenta a maior porcentagem de ocorrência de usos das imagens disponíveis, com o uso de 6 imagens distintas em 8 páginas da Wikipédia em português e em 10 artigos de outros idiomas, resultando na ocorrência do total de oito reúsos de imagens.

Tabela 1 – Reuso de imagens nos projetos da Fundação Wikimedia

Coleção	Arquivos na categoria	Wikipédia português	Wikipédia em outros idiomas	Wikibooks	Wikiquote	Wikidata
Museu da Inconfidência	7	8	10	0	0	0
Museu da República	9	33	75	1	1	0
Museu Nacional de Belas Artes	17	2	4	0	0	1
Museus Castro Maya	46	12	25	5	0	3
Museu Casa de Benjamin Constant	31	4	0	0	0	0
Museu Regional de São João del-Rei[4]	6	0	0	0	0	0
Museu Victor Meirelles	4	5	12	0	0	3
Museu Histórico Nacional	471	13	19	0	0	4
Museu Imperial	38	7	52	0	0	0

Fonte: Dados da pesquisa, 2020[5].

Tabela 2 – Quantidade de imagens reutilizadas nos projetos da Fundação Wikimedia

Coleção	Arquivos na categoria	Total de imagens distintas	Total de uso de imagens	Imagens da categoria (%)
Museu da Inconfidência	7	6	18	85,71
Museu da República	9	4	110	44,44
Museu Nacional de Belas Artes	17	7	7	17,65
Museus Castro Maya	46	17	45	36,96
Museu Casa de Benjamin Constant	31	4	4	12,9
Museu Regional de São João del-Rei	6	0	0	0
Museu Victor Meirelles	4	3	20	75
Museu Histórico Nacional	471	12	36	2,55
Museu Imperial	38	8	59	21,05

Fonte: Dados da pesquisa, 2020.

[4] Apesar de identificada a categoria com 6 mídias referentes à coleção ao Museu de São João del-Rei, nenhuma das ferramentas utilizadas retornou dados sobre ela.

[5] Os dados da pesquisa podem ser acessados em: https://docs.google.com/spreadsheets/d/1BNYn2WKGGqOT0UvC_Nla9OgciK3I8V_6NRfg5ckE4-8/edit?usp=sharing

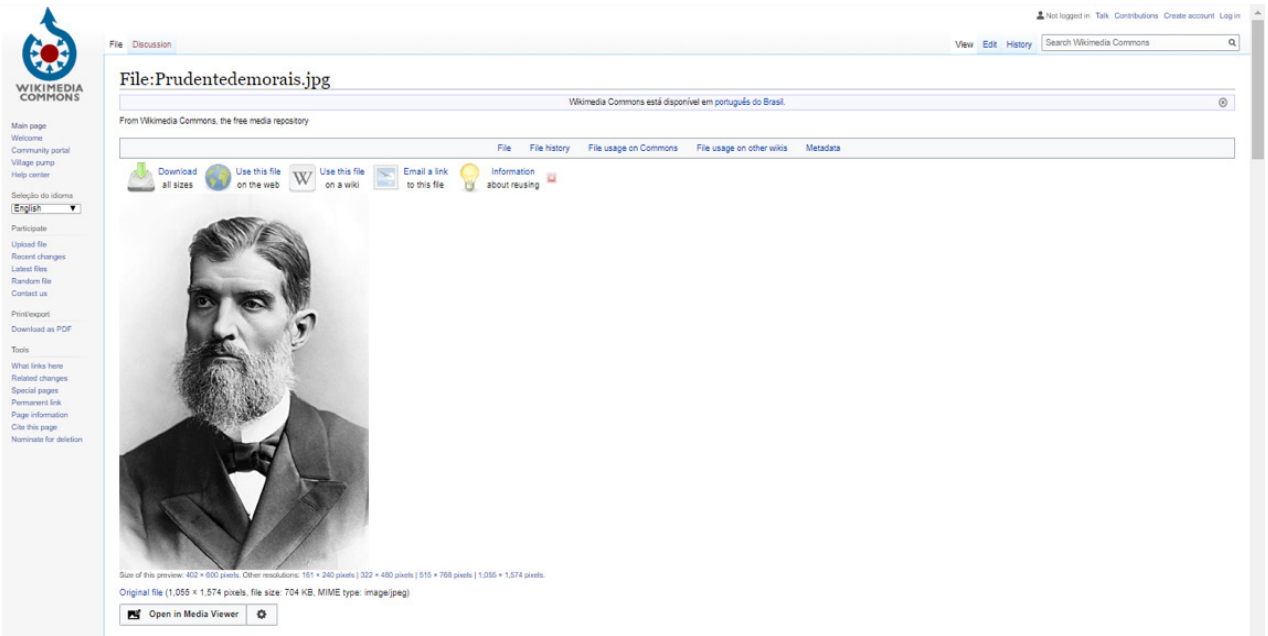
A REUTILIZAÇÃO DAS IMAGENS DAS CATEGORIAS REFERENTE ÀS COLEÇÕES DE MUSEUS NA WIKIPÉDIA

A Wikipédia é um projeto de enciclopédia on-line de edição coletiva e colaborativa da Fundação Wikimedia. Devido à sua proposta multilíngue, atualmente existem versões da Wikipédia em 303 idiomas. Cada idioma se mostra como um projeto Wikipédia independente do outro, ou seja, os conteúdos não são automaticamente traduzidos e reproduzidos, sendo originalmente criados e geridos pelos usuários de cada comunidade linguística e apresentam seu subdomínio. A Wikipédia, em seus variados idiomas, tem seus conteúdos baseados na licença aberta Creative Commons Attribution-ShareAlike.

Atualmente[6] a Wikipédia lusófona alcança a marca de 1.029.760 artigos produzidos, 6.110 usuários ativos (WIKIPEDIA, 2019).

Para entender melhor como se dá a reutilização dessas imagens tanto em páginas da Wikipédia lusófona, quanto em outros idiomas, buscamos coletar dados sobre a imagem da categoria referente ao Museu da República que foi reutilizada com maior frequência[7]. Por meio da ferramenta GLAMorous verificamos que a imagem com maior utilização é uma fotografia do ex-presidente brasileiro Prudente de Morais. Esse arquivo intitulado "Prudentedemorais.jpg" aparece com 95 ocorrências de reutilização.

Figura 1 – Retrato de Prudente de Morais na Wikimedia Commons



Fonte: Wikimedia Commons, 2020[8].

[6] Dados coletados em abril de 2020.
[7] Em abril de 2020.
[8] Disponível em: <https://commons.wikimedia.org/wiki/File:Prudentedemorais.jpg>. Acesso em abr. 2020.

Ao observar a imagem no contexto da plataforma Wikimedia Commons, é possível verificar que a página da imagem oferece meios de download e compartilhamento da imagem, e também exibe uma série de seções com informações da mídia. A seção intitulada "Informações do arquivo" fornece uma série de informações sobre a imagem como "Descrição", "Data", "Permissão"[9] e "Outras versões".

Além disso, podem ser observadas outras seções como "Histórico do arquivo", que mostra o histórico de substituição de arquivos no Wikimedia Commons, a seção "Uso de arquivos no Commons", que dá informações sobre o uso de versões no Commons, a seção "Uso de arquivos em outras wikis", que gera uma lista de projetos wiki nos quais a imagem está sendo utilizada, e por fim a seção "Metadado", que reúne metadados e dados específicos da mídia reprodutora da imagem, que nesse caso são específicos à fotografia que resultou na imagem digitalizada.

Ao consultar a listagem na seção "Uso de arquivos de outras wikis", é possível observar em quais outros idiomas de Wikipédias o retrato de Prudente de Morais está sendo utilizado e em quais páginas específicas. Essas informações também podem ser obtidas por meio de consultas à ferramenta GLAMorous.

Assim foi possível verificar a presença do retrato de Prudente de Morais em artigos de Wikipédias em 44 idiomas distintos e também em uma página da Wikiquote em português. Também foi possível verificar que a mesma imagem está presente em 30 artigos da Wikipédia em português. No quadro 2 é possível verificar a listagem desses artigos.

Quadro 2 − Lista de artigos que reutilizam a imagem o retrato de Prudente de Morais

pt.wikipédia	Prudente de Morais Rodrigues Alves Presidente Prudente Lista de eleições presidenciais no Brasil Presidente do Brasil Bernardino José de Campos Júnior Manuel José Alves Barbosa Eleição presidencial no Brasil em 1891 Amaro Cavalcanti Manuel Vitorino Joaquim Murtinho Antônio Olinto dos Santos Pires Antônio Gonçalves Ferreira Dionísio Evangelista de Castro Cerqueira Alberto Torres Carlos Machado de Bittencourt Carlos Augusto de Carvalho Lista de presidentes do Senado Federal do Brasil João Tomás de Cantuária Jerônimo Rodrigues de Morais Jardim Sebastião Eurico Gonçalves de Lacerda Francisco de Paula Argolo Bernardo Vasques Elisiário José Barbosa Eleição presidencial no Brasil em 1894 Lista de presidentes do Brasil Governo Prudente de Morais Lista de senadores do Brasil da 23.ª legislatura Lista de senadores do Brasil da 22.ª legislatura

Fonte: Dados da pesquisa, 2020.

Percebe-se que a imagem é reutilizada sobretudo para ilustrar verbetes de personagens históricos que tiveram relação com o ex-presidente ou eventos importantes relacionados a ele. Além disso, aparecem na lista alguns eventos históricos, listas de senadores em legislaturas específicas, eleições presidenciais e o próprio verbete relativo ao seu governo. A imagem do acervo do museu é colocada em contexto a partir da textualidade dos verbetes e das relações simbólicas que estabelece com outras imagens e elementos textuais. Estudar essas relações disponibiliza uma camada informacional de compreensão do reúso da informação a ser desenvolvida em pesquisas posteriores, mas é algo a se destacar como potencial para a área.

[9] Permissões de reúso do artigo, especificando as licenças atribuídas.

A REUTILIZAÇÃO DE IMAGENS NO REPOSITÓRIO DE DADOS ESTRUTURADOS WIKIDATA

Outro aspecto interessante a ser verificado é a reutilização das imagens de coleções do universo de museus e instituições culturais na ilustração de itens do Wikidata. O Wikidata é o projeto de repositório de dados estruturados da Fundação Wikimedia e tem "como objetivo a criação de uma base de conhecimento livre sobre o mundo que pode ser lida e editada por humanos e máquinas" (WIKIDATA, 2019). Segundo dados coletados em abril de 2020 na página do projeto, o Wikidata registrava 83.779.497 milhões de itens em seu banco, 26.160 usuários ativos no mês da coleta e 308 *bots* que auxiliam nas edições. Os dados publicados no Wikidata também são disponibilizados de forma aberta sob a licença Creative Commons Public Domain Dedication 1.0.

As regras de criação e gerenciamento do conteúdo são de responsabilidade da comunidade de editores, que devem seguir os mesmos princípios e políticas de colaboração que norteiam os demais projetos Wikimedia.

Para visualizar e entender a reutilização de imagens do Wikimedia Commons no Wikidata, selecionamos para a análise a coleção que apresentou, na tabela 1, o número maior de ocorrências de uso de imagens no Wikidata, que foi o caso do Museu Histórico Nacional. Usando a ferramenta GLAMorous e explorando os resultados detalhados retornados por meio de busca com o nome da categoria, foi possível verificar quais imagens foram utilizadas nas 4 ocorrências detectadas e também o código identificador do item no Wikidata, como pode ser visto no quadro 3.

Quadro 3 – Imagens reutilizadas na Wikidata

Nome do arquivo	ID Wikidata
Conselheiro Francisco de Carvalho Soares Brandão, da coleção Museu Histórico Nacional.jpg	Q62091980
Beatriz Reinall, da coleção Museu Histórico Nacional.jpg	Q62091982
Caminho na floresta, da coleção Museu Histórico Nacional.jpg	Q62091983
Viscondessa do Bom Conselho, da coleção Museu Histórico Nacional.jpg	Q62091547

Fonte: Dados da pesquisa, 2020.

Figura 2 – Retrato de Prudente de Morais na Wikimedia Commons.

Fonte: Wikimedia Commons, 2020[10].

[10] Disponível em: < https://www.wikidata.org/wiki/Q62091980>. Acesso em 15 de abr. 2020.

Para visualizar como se dá essa reutilização no contexto da plataforma Wikidata, realizamos buscas na plataforma por meio do identificador do arquivo da imagem intitulada "Conselheiro Francisco de Carvalho Soares Brandão, da coleção Museu Histórico Nacional.jpg". Desse modo foi possível observar que a mídia foi inserida na propriedade "imagem" do item no Wikidata, que apresenta dados estruturados sobre a obra cuja original analógica está sob guarda no acervo do Museu Histórico Nacional.

Ao verificar a reutilização do arquivo intitulado "Beatriz Reinall, da coleção Museu Histórico Nacional.jpg", foi possível identificar que essa imagem também ilustra um item do Wikidata que apresenta dados estruturados sobre a obra de arte retratada. A mesma situação se ocorre no caso das outras duas imagens identificadas que também ilustram itens sobre as obras originais. Com essas informações, é possível inferir que foi realizado um trabalho coordenado de descrição de obras do Museu Histórico Nacional na Wikidata. Em uma investigação realizada em 2019, os autores Carmo e Martins (2019, p. 12) identificaram que o Museu Histórico Nacional representava o segundo maior acervo oriundo de instituições brasileiras com itens de pinturas descritas no Wikidata, reunindo 503 itens. Esses dados indicam que a instituição estabeleceu em algum momento, ou ainda estabelece, práticas que adotam as plataformas da Fundação Wikimedia como meios estratégicos de disseminação de seus acervos.

A VISUALIZAÇÃO DE ARQUIVOS DE MÍDIA DO WIKIMEDIA COMMONS

Utilizando a ferramenta GLAMorous 2, foi possível obter dados de visualização dos arquivos de mídias disponibilizados na Wikimedia Commons que estão nas categorias referentes às coleções de museus geridos pelo Ibram. Essas visualizações são contabilizadas com base nos dados de acesso a páginas de projetos da Fundação Wikimedia, como o Wikidata, Wikibooks e Wikiquote, e também das Wikipédias dos mais variados idiomas que exibem imagens do Wikimedia Commons reutilizadas em suas páginas. O recorte utilizado alcança dados de visualização relativos aos 12 meses do ano de 2019, de janeiro a dezembro. Na tabela 3 podemos verificar o total de visualizações de arquivos de mídia no ano de 2019, assim como a média mensal. Também foi possível observar o total de visualizações somente na Wikipédia em português, assim como o total de visualizações em outras Wikipédias e projetos.

Com base nos dados coletados, é possível observar alguns números que nos chamam especial atenção. Tomamos como exemplo os dados de visualização da coleção referente ao Museu da República, que com apenas 9 arquivos na categoria foi a que gerou o número de visualizações de arquivos de mídias mais expressivo no ano de 2019, alcançando a marca de 4.147.494 de visualizações.

Tabela 3 – Reúso de imagens nos projetos da Fundação Wikimedia

Coleção	Total de visitas - 2019	Média mensal	Total de visualizações pt wikipédia	Total de visualizações outras wikis
Museu da Inconfidência	675.728	56.310,67	635.523	40.205
Museu da República	4.147.494	345.624,50	3.225.730	921.764
Museu Nacional de Belas Artes	21.873	1.822,75	20.904	969
Museus Castro Maya	882.608	73.550,67	795.007	87.601
Museu Casa de Benjamin Constant	6.216	518,00	6.216	0
Museu Regional de São João del-Rei	0	-	0	0
Museu Victor Meirelles	69.089	5.757,42	44.791	24.298
Museu Histórico Nacional	578.668	48.222,33	45.643	45.643
Museu Imperial	650.276	54.189,67	174.333	475.943

Fonte: Dados da pesquisa, 2020.

Ao recorrer à opção "Detalhes de uso do arquivo" na ferramenta GLAMorous2, é possível observar que a imagem identificada como "Prudentedemorais.jpg"[11], já anteriormente identificada como a mídia que mais apresenta reúsos dentro da coleção, revela-se sozinha como a responsável por 3.096.852 do total de visualizações das mídias que estão na categoria da coleção do museu. Seguida da imagem intitulada "GetulioVargasPijamaRevolver.jpg", que registrou no ano de 2019 o total de 992.687 visualizações. Também foi possível constatar que 3.225.730 visualizações das mídias da coleção do Museu da República foram somente em Wikipédias em português, e 675.728 em Wikipédias de outros idiomas e em outros projetos da Wikimedia. A título de ilustração, para que se possa problematizar o fenômeno social em questão, no primeiro semestre do ano de 2019 o Museu da República teve em torno de 105.000 visitantes presenciais, segundo dados fornecidos pelo próprio Instituto Brasileiro de Museus (BRASIL, 2019). Se tomarmos por partido a comparação com os números digitais, temos que o número de pessoas que visualizaram objetos do acervo presencialmente representa em torno de 2,5% dos que visualizaram os objetos digitais. Sem dúvida, e é fundamental destacar esse ponto, não se quer com isso fazer uma equivalência das experiências, sabendo que as visitas presenciais são mediadas de diversos outros elementos culturais que precisam ser reconhecidos e que possuem efeitos importantes na apropriação. Mas compreender esse espaço de socialização e sua dimensão simbólica como estratégia complementar para o campo dos museus é algo de enorme importância nos tempos atuais. Há um espaço de trabalho que pode ser aprimorado e que contém largo potencial de socialização e produção de valor social pelo museu.

Outro dado interessante foi a quantidade de visualizações de mídias referente a categoria do Museu Imperial, que apresentou maior número de visualizações em Wikipédias de outros idiomas, com 475.943 visualizações, do que na Wikipédia em português, com 174.333 visualizações. Ao verificar esses dados na aba da ferramenta chamada "Uso global do arquivo", foi possível constatar que só a Wikipédia em inglês indica o total de 319.974 visualizações e faz 14 usos de 7 diferentes mídias da coleção do Museu Imperial em suas páginas[12], no ano 2019.

Já as mídias da categoria referente ao Museu Benjamin Constant chamam atenção por não apontarem visualizações em outros projetos da Wikimedia ou em Wikipédias de outros idiomas, somente na Wikipédia em português com 6.216 visualizações no ano de 2019, representando a coleção que menos obteve visualizações dentro do universo investigado.

CONCLUSÃO

O presente artigo trouxe resultados de uma investigação que buscou explorar possibilidades de coleta de dados que permitissem a realização de análises webométricas que permitem mensurar aspectos relacionados ao acesso e à reutilização de mídias de acervos brasileiros disponíveis na internet, por meio do repositório de mídias Wikimedia Commons. Assim, verificamos a presença de coleções de nove dos 30 museus sob responsabilidade do Ibram entre as categorias do Wikimedia Commons. Identificadas essas categorias e com o auxílio das ferramentas apresentadas, obtivemos e descrevemos números relativos à reutilização dessas mídias tanto em páginas da Wikipédia em português, das Wikipédias em outros idiomas, quanto em outros projetos da Fundação Wikimedia.

[11] A ferramenta GLAMorous 2 apresenta também a estatística de uso da imagem "Prudentemorais,jpg", que indica 106 usos. Ao comparar com os dados obtidos pela GLAMorous, foi possível observar uma discrepância, já que ela nos apresenta um número diferente de usos para essa imagem, neste caso 95 usos. Ao verificar os dados e mais especificamente a lista que especifica as páginas, percebemos que a ferramenta GLAMorous 2 conta todas as páginas dos projetos, incluindo páginas de usuários que não são artigos, enquanto a GLAMorous mostra somente os usos em artigos.

[12] Esses dados podem ser verificados no link de consulta: <https://tools.wmflabs.org/glamtools/glamorous/?mode=category&text=Collections%20of%20the%20Museu%20Imperial&pageviews=1&month=2019-01%7C2019-12> Acesso em 15 de abril, 2020.

Assim foi possível observar que o número de mídias disponíveis não afeta diretamente o número de reutilizações ou de visualizações, já que o Museu da República se destacou tanto em número de reúsos quanto em números de visualizações, com apenas 9 arquivos na categoria. Apesar de o Museu Histórico Nacional conter o maior número de mídias, 471 arquivos, não obteve números relevantes nos dados de reúso ou de visualização deles. Também foi possível observar que a mesma imagem de determinada coleção pode ser utilizada diversas vezes, gerando uma porcentagem baixa do total de uso dos arquivos. Outra situação que se repetiu foi a baixa reutilização de mídias para ilustrar outros projetos da Fundação Wikimedia, como o Wikidata. A presença de mídias e a quantidade de visualizações nas Wikipédias de diferentes idiomas foram dados que nos chamaram atenção, com ênfase no Museu da República, e mais especificamente no caso do retrato do ex-presidente brasileiro Prudente de Morais, que aparece ilustrando 30 artigos da Wikipédia em português e 45 artigos de Wikipédias de outros idiomas. Outro dado curioso a ser mencionado foi o fato de o Museu Imperial apresentar o maior número de visualizações na Wikipédia em inglês do que na Wikipédia em português. Por fim, outro dado relevante foi o número de visualizações obtidos pelos arquivos na categoria referente a coleções do Museu da República, que concentrou em apenas uma imagem mais de 3 milhões de visualizações, somente no ano de 2019. Durante a investigação foi possível perceber evidências que apontam para a expressividade do público digital em relação ao público presencial, deixando elementos que podem ser explorados em pesquisas posteriores, procurando qualificar e entender que público é esse, assim como os impactos que esse tipo de visitação pode gerar para os museus e como isso pode refletir nas práticas de seus profissionais.

Por meio dos dados obtidos e das análises realizadas foi possível observar que a plataforma Wikimedia Commons, assim como outras plataformas da Fundação Wikimedia, como a Wikipédia e Wikidata, são espaços sociotécnicos que além de se configurar como meios potenciais de socialização de acervos, também se configuram como fonte de informação valiosa sobre a manipulação e uso das informações disponibilizadas por meio deles. Acreditamos que a obtenção desse tipo de dado pode ajudar instituições culturais guardiãs de acervos, como os museus, a conhecer e a compreender o que acontece com seus acervos digitais uma vez que são disponibilizados nesses espaços, fornecendo importantes informações para a gestão das coleções de seu acervo digital, principalmente em relação ao acesso e reapropriação desses acervos por parte dos usuários.

REFERÊNCIAS

BJÖRNEBORN, L.; INGWERSEN, P. Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, v. 55, n. 14, p. 1216–1227, dez. 2004. DOI 10.1002/asi.20077. Disponível em: http://doi.wiley.com/10.1002/asi.20077. Acesso em: 12 mar. 2021.

BRASIL. MINISTÉRIO DA CIDADANIA. *Plano Nacional de Cultura. Museus brasileiros apresentam aumento no número de visitantes*. 19 ago. 2019. Disponível em: http://pnc.cultura.gov.br/2019/08/19/museus-brasileiros-apresentam-aumento-no-numero-de-visitantes/. Acesso em: 30 abr. 2020.

CARMO, D. do; MARTINS, D. L. A presença dos museus brasileiros na ecologia informacional da Fundação Wikimedia: estudo de caso do projeto Sum of All Paintings. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 20., 2019. *Anais* [...]. Florianópolis: UFSC, 2019. v. 20, p. 1–20. Disponível em: https://conferencias.ufsc.br/index.php/enancib/2019/paper/view/909/679. Acesso em: 12 mar. 2021.

INSTITUTO BRASILEIRO DE MUSEUS. *Museus IBRAM*. 2020. Disponível em: https://www.museus.gov.br/os-museus/museus-IBRAM/. Acesso em: 5 abr. 2020.

MANSKE, M. *GLAMorous*. 2020a. Disponível em: https://tools.wmflabs.org/glamtools/glamorous.php. Acesso em: 18 abr. 2020.

MANSKE, M. *GLAMorous 2*. 2020b. Disponível em: https://tools.wmflabs.org/glamtools/glamorous/. Acesso em: 18 abr. 2020.

STINSON, A. D.; WYATT, L.; FAUCONNIER, S. Stepping Beyond Libraries : the Changing Orientation in Global GLAM-Wiki. *JLIS*, n. 9, 2018. DOI 10.4403/jlis.it-12480. Disponível em: https://doi.org/10.4403/jlis.it-12480. Acesso em: 12 mar. 2021.

THELWALL, M. Introduction to Webometrics: Quantitative Web Research for the Social Sciences. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, v. 1, n. 1, p. 1–116, jan. 2009. DOI 10.2200/S00176ED1V01Y200903ICR004. Disponível em: http://www.morganclaypool.com/doi/abs/10.2200/S00176ED1V01Y200903ICR004. Acesso em: 12 mar. 2021.

VANTI, N. A. P. Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento. *Ciência da Informação*, v. 31, n. 2, p. 369–379, ago. 2002. DOI 10.1590/S0100-19652002000200016. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652002000200016&lng=pt&tlng=pt. Acesso em: 12 mar. 2021.

VILLAESPESA, E.; NAVARRETE, T. Museum Collections on Wikipedia: Opening Up to Open Data Initiatives. *In*: MUSEWEB, 14 jan. 2019. *Proceedings* [...]. Boston: [*s. n.*], 14 jan. 2019. Disponível em: https://mw19.mwconf.org/paper/museum-collections-on-wikipedia-opening-up-to-open-data-initiatives/. Acesso em: 12 mar. 2021.

WIKIMEDIA COMMONS. *Commons: Welcome*. 2020a. Disponível em: https://commons.wikimedia.org/wiki/Commons:Welcome. Acesso em: 15 abr. 2020.

WIKIMEDIA COMMONS. *Special: Statistics*. 2020b. Disponível em: https://commons.wikimedia.org/wiki/Special:Statistics. Acesso em: 18 abr. 2020.

WIKIMEDIA OUTREACH. *GLAM/Resources/Tools*. 2020. Disponível em: https://outreach.wikimedia.org/wiki/GLAM/Resources/Tools. Acesso em: 20 mar. 2020.

WIKIPEDIA. *Página Principal*. 2019. Disponível em: https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:P%C3%A1gina_principal. Acesso em: 18 out. 2020.

ZEINSTRA, M. *Report on Requirements for Usage and Reuse Statistics for GLAM Content*. [*S. l.*]: Wikimedia, 31 maio 2013. Disponível em: https://www.kl.nl/en/publications/report-on-requirements-for-usage-and-reuse-statistics-for-glam-co/. Acesso em: 20 set. 2019.

Google Dataset Search: Visão geral e perspectivas para indexação e disponibilização de conjuntos de dados científicos abertos

Adilson Luiz Pinto
Pós-Doutorado pelo Institut de Recherche en Sciences de l´Information et de la Communication (IRSIC) - França. Doutor em Documentação pela Universidad Carlos III de Madrid (UC3M) - Espanha. Professor da Universidade Federal de Santa Catarina (UFSC) - Florianópolis, SC - Brasil.
http://lattes.cnpq.br/4767432940301118
E-mail: adilson.pinto@ufsc.br

Eduardo Diniz Amaral
Doutorando em Ciência da Informação pela Universidade Federal de Santa Catarina (UFSC) – SC - Brasil. Mestre em Biotecnologia pela Universidade Estadual de Montes Claros (Unimontes) - Montes Claros, MG - Brasil. Professor da Universidade Estadual de Montes Claros (Unimontes) - Montes Claros, MG - Brasil.
http://lattes.cnpq.br/1241483438206633
E-mail: eduardo.diniz@unimontes.br

Submetido em: 25/11/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021.

RESUMO

Com o intuito de colaborar com a produção científica na área de ciência de dados, especificamente em ferramentas de armazenamento e recuperação de conjuntos de dados pela internet, este artigo tem como propósito obter uma visão geral do funcionamento, padrões e perspectivas sobre a ferramenta *Google Dataset Search* –lançada em 2018 com a proposta de identificar, indexar e disponibilizar pela internet *datasets* (conjuntos massivos de dados) - instrumentos salutares para a comunidade científica. A metodologia utilizada foi descritiva, de caráter exploratório e bibliográfica sobre o tema. Foi realizado levantamento bibliográfico sobre a plataforma, identificando funcionamento interno, padrões, diretrizes, formatos e instituições de padronização que norteiam a plataforma, além de estatísticas atuais de dados indexados. Em seguida, foram executados testes práticos de utilização, usabilidade e funcionamento da ferramenta, conforme documentação disponível. Os resultados obtidos mostraram uma plataforma promissora, com índice satisfatório de usabilidade, alinhada com padrões internacionais de interoperabilidade de dados e com volumes consideráveis de *datasets* já disponíveis, em sua grande maioria no idioma inglês. Observou-se ainda, após os testes, que já existem diversos repositórios brasileiros de dados indexados pelo *Google Dataset Search*. Entretanto, alguns deles, mesmo adotando iguais padrões de metadados desta ferramenta, ainda não estão disponíveis. A conclusão é que se trata de um sistema criado pela Google, com alta capacidade de rastreamento, identificação, indexação, interoperação e disponibilização de conjuntos de dados disponíveis na internet utilizando padrões internacionais e, por isso, apresenta expressivo potencial. Este trabalho contribui para a grande área que está inserido reduzindo a escassez de publicações científicas acerca de ferramentas de disponibilização de conjuntos de dados, especificamente sobre o funcionamento, protocolos, mecanismos e interface da ferramenta em questão.

Palavras-chave: Conjuntos de dados. Interoperabilidade. Acesso aberto. Padrões de metadados. *Google Dataset Search.*

Google Dataset Search: Overview and perspectives for indexing and availability of open scientific datasets

ABSTRACT

In order to collaborate with scientific production in the field of data science, specifically in tools for storage and retrieval of data sets over the internet, this article aims to obtain an overview of the functioning, standards and perspectives on the Google Dataset Search tool - launched in 2018 with the proposal of identifying, indexing and making available internet datasets (massive sets of data) - essential instruments for the scientific community. The methodology used was descriptive, exploratory and bibliographic. A bibliographic survey was carried out on the platform, identifying internal functioning, standards, guidelines, formats and standardization institutions that guide the platform, in addition to current statistics of indexed data. Then, practical tests of use, usability and operation of the tool were performed, according to available documentation. The results obtained showed a promising platform, with a satisfactory usability score, aligned with international data interoperability standards and with considerable volumes of datasets already available, mostly in the English language. It was also observed, after the tests, that there are already several brazilian data repositories indexed by Google Dataset Search. However, some of them, even adopting the same metadata standards as this tool, are not yet available. The conclusion is that it is a system created by Google, with a high capacity for tracking, identification, indexing, interoperation and making available data sets available on the internet using international standards and, therefore, has significant potential. This work contributes to the large area that is inserted, reducing the scarcity of scientific publications on tools for making data sets available, specifically on the functioning, protocols, mechanisms and interface of this current tool.

***Keywords**: Data sets. Interoperability. Open access. Metadata standards. Google Dataset Search.*

Google Dataset Search: descripción general y perspectivas para indexar y poner a disposición conjuntos de datos científicos abiertos

RESUMEN

Para colaborar con la producción científica en el campo de la ciencia de datos, específicamente en herramientas para el almacenamiento y recuperación de conjuntos de datos a través de Internet, este artículo tiene como objetivo obtener una descripción general del funcionamiento, los estándares y las perspectivas de la herramienta Google Dataset Search, lanzada en 2018 con la propuesta de identificar, indexar y poner a disposición conjuntos de datos de Internet (conjuntos masivos de datos), instrumentos saludables para la comunidad científica. La metodología utilizada fue descriptiva, exploratoria y bibliográfica sobre el tema. Se realizó un relevamiento bibliográfico, identificando funcionamiento interno, estándares, lineamientos, formatos e instituciones de estandarización que orientan la plataforma, además de estadísticas actuales de datos indexados. A continuación, se realizaron pruebas prácticas de uso, usabilidad y funcionamiento de la herramienta, según documentación disponible. Los resultados obtenidos mostraron una plataforma prometedora, con un índice de usabilidad satisfactorio, alineada con los estándares internacionales de interoperabilidad de datos y con volúmenes considerables de conjuntos de datos ya disponibles, en su mayoría en idioma inglés. También se observó, después de las pruebas, que ya existen varios repositorios de datos brasileños indexados por Google Dataset Search. Sin embargo, algunos de ellos, incluso adoptando los mismos estándares de metadatos que esta herramienta, aún no están disponibles. La conclusión es que se trata de un sistema creado por Google, con una alta capacidad de seguimiento, identificación, indexación, interoperación y puesta a disposición de conjuntos de datos en Internet utilizando estándares internacionales y, por tanto, tiene un potencial significativo. Este trabajo contribuye a la gran área que se inserta, reduciendo la escasez de publicaciones científicas sobre herramientas para la puesta a disposición de conjuntos de datos, específicamente sobre el funcionamiento, protocolos, mecanismos e interfaz de la herramienta en cuestión.

***Palabras clave**: Conjuntos de datos. Interoperabilidad. Acceso abierto. Estándares de metadatos. Google Dataset Search.*

INTRODUÇÃO

De acordo com Gavron e Canto (2017), o acesso aberto à informação científica exerce importante influência no desenvolvimento da ciência, pois, por meio do acesso aberto é possível conhecer o que está sendo realizado pelos pesquisadores em todas as partes do globo. Quanto mais atualizado, mais relevante será para os pesquisadores, promovendo um melhor diálogo e intercâmbio informacional entre eles. Neste contexto, os *datasets* (ou conjunto de dados) representam alta relevância. Trata-se de coleções brutas de dados organizados sobre um tema ou contexto específico, geralmente dispostos em colunas (como atributos) e linhas como dados individuais, elementos ou unidades, nos mais diversos formatos (planilhas, arquivos texto, listas, tabelas etc).

No que tange o universo científico, os conjuntos de dados coletados, organizados e armazenados em experimentos, por exemplo, fazem parte do que chamamos de cauda longa de pesquisa, e são fundamentais para replicação, comprovação e novas análises destes. São informações valiosas que, se bem trabalhadas, podem gerar novos caminhos para pesquisas científicas.

Existem disponíveis na internet milhares de repositórios de dados, provendo acesso a milhões de *datasets* (NOY, 2020). Assim, dada a importância destes repositórios, os esforços de instituições nacionais e internacionais para indexar e organizar conjuntos de dados abertos ao público é cada vez maior.

No Brasil, por exemplo, temos, dentre estas iniciativas, o Portal Brasileiro de Dados Abertos, plataforma disponibilizada pelo governo brasileiro para que todos possam encontrar e utilizar dados e informações públicas (BRASIL, 2019). No cenário internacional a empresa *Google* – uma das maiores empresas da indústria informacional eletrônica contemporânea – lançou, em setembro de 2018, a ferramenta *Google Dataset Search* (referenciado pela empresa com a sigla GOODS), que se propõe a localizar e indexar *datasets* disponíveis na internet, promovendo a descoberta e catalogação destes repositórios (através de *harvesting*, inteligência artificial, big data e outras tecnologias de dados) bem como a disponibilização destes através de interfaces de consulta, desde que estejam de acordo com os padrões de dados e metadados estabelecidos internacionalmente.

Assim, dada a relevância das ferramentas disponibilizadas pela referida empresa, bem como a sua notável predominância no mercado informacional, o escopo e objeto de estudo deste trabalho orbita o buscador *Google Dataset Search*.

Configura-se como objetivo principal obter uma visão do GOODS, identificando aspectos funcionais e técnicos, bem como dos padrões de dados utilizados e por fim perspectivas acerca da ferramenta. Como objetivos específicos, lista-se:

a) Mapear estrutura de funcionamento e funções da ferramenta através da aplicação de testes;

b) Identificar procedimentos e padrões de interoperabilidade entre *datasets* e a plataforma;

c) Realizar buscas de testes em sites de domínios .br, conforme metodologias específicas e;

d) Conjecturar sobre as perspectivas da ferramenta e seus impactos informacionais para a comunidade científica.

METODOLOGIA

De acordo com Gerhardt e Silveira (2009), esta pesquisa caracteriza-se como natureza aplicada, descritiva e de caráter exploratório. É também bibliográfica, pois fez uso de artigos, manuais e outros documentos a respeito ferramenta, disponibilizados pela própria empresa, disponíveis no site oficial, https://developers.google.com, e também por outros autores especialistas nesta temática.

Inicialmente, foi realizado um levantamento bibliográfico sobre a ferramenta, com o intuito de descrever seu funcionamento, estrutura tecnológica e padrões de interoperabilidade. Tal estudo foi realizado em meados do mês de agosto de 2019 através do buscador google.com e scholar.google.com, aplicando o termo "*google dataset Search*", e atualizado em setembro de 2020. Após o levantamento bibliográfico iniciou-se o processo de testes do GOODS.

Como o sistema funciona em ambiente web, os testes de interface e usabilidade foram realizados aplicando-se as 10 heurísticas propostas por Jakob Nielsen (ROSA; VERAS, 2013), específicas para este fim. São elas: Visibilidade do status do sistema; Compatibilidade entre o sistema e o mundo real; Controle e liberdade para o usuário; Consistência e padronização; Prevenção de erros; Reconhecimento ao invés de memorização; Eficiência e flexibilidade de uso; Estética e Design minimalista; Ajude os usuários a reconhecerem, diagnosticarem e recuperarem-se de erros; e Ajuda e documentação. Quanto às funcionalidades, aplicou-se testes funcionais de unidade no método caixa-preta - técnica de teste software em que, de acordo com Myers, Sandler e Badgett (2012), o objeto a ser testado é abordado como se fosse uma caixa-preta, ou seja, dados de entrada são inseridos, o teste é executado e o resultado do processamento obtido é comparado ao resultado esperado previamente conhecido, considerando-se o desconhecimento do funcionamento interno do sistema. Haverá sucesso na aplicação do teste se o resultado obtido for de acordo com o resultado esperado.

Por fim, com base nos levantamentos e análises práticas, além de informações e opiniões sobre a ferramenta, recursos das últimas versões e notícias, foram realizadas conjecturas sobre as perspectivas do GOODS para a organização e disponibilização dos conjuntos de dados para a comunidade científica.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

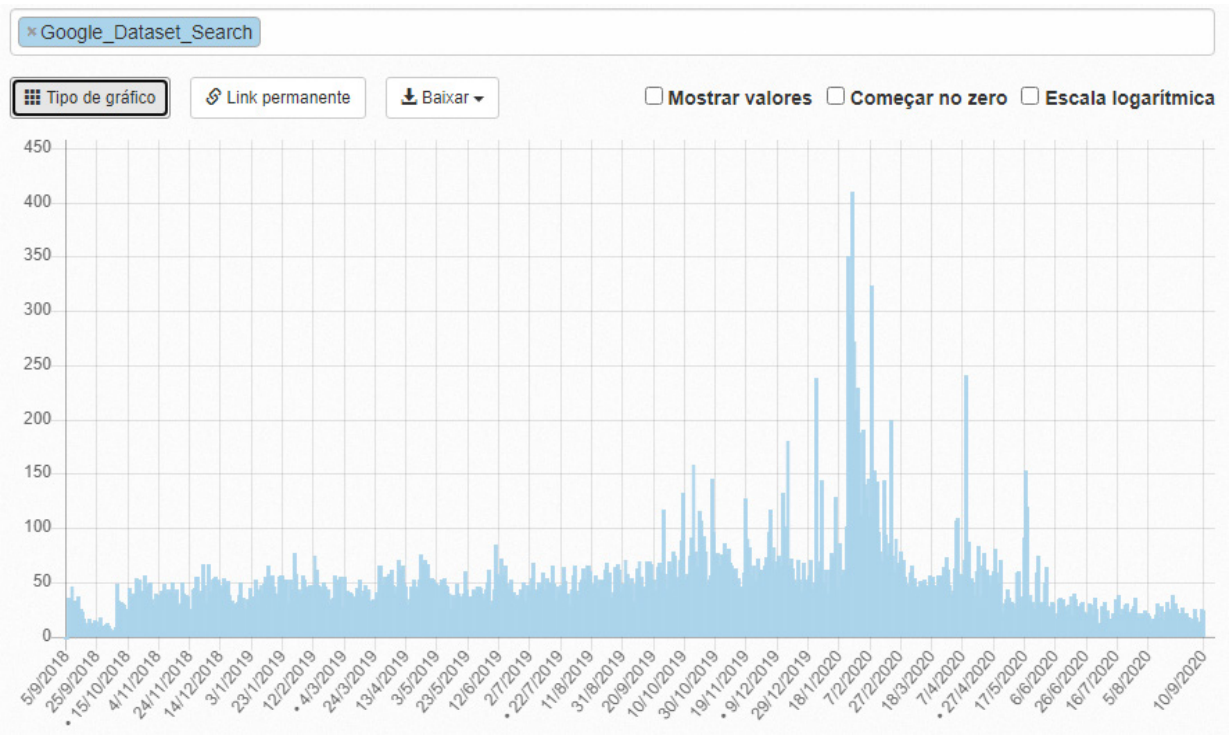
Serão apresentados a seguir os dados obtidos através dos levantamentos bibliográficos e, em seguida, os resultados dos testes realizados.

SOBRE O *GOOGLE DATASET SEARCH*

Na data da submissão deste trabalho foram encontrados aproximadamente 37.300 resultados para "*google dataset search*" no buscador da *Google*, incluindo notícias e postagens em fóruns técnicos especialistas. Efetuando a mesma busca no Google Acadêmico (scholar.google.com) foram localizados aproximadamente 335 resultados.

Assim, em comparação com a busca convencional, poucos artigos e documentos científicos foram encontrados, o que aparenta ser justificado pelo relativo espaço de tempo entre a data de lançamento da ferramenta e a data atual. Grande parte destes foram escritos por desenvolvedores, funcionários e cientistas da própria *Google*. Lançado em 2018, em 23 de janeiro de 2020 o *Google Dataset Search* (GOODS) deixou de ser beta, passando a integrar o rol de produtos oficiais da empresa, oferecendo cerca de 25 milhões de *datasets* indexados, além de novas funcionalidades (NOY, 2020). A página no idioma inglês sobre esta ferramenta na enciclopédia *Wikipedia* aponta cerca de 38.200 visualizações, com auge na ocasião do evento de lançamento oficial, conforme apontado na figura 1.

Figura 1 – Visualização de acessos da página do Google Dataset Search na Wikipedia

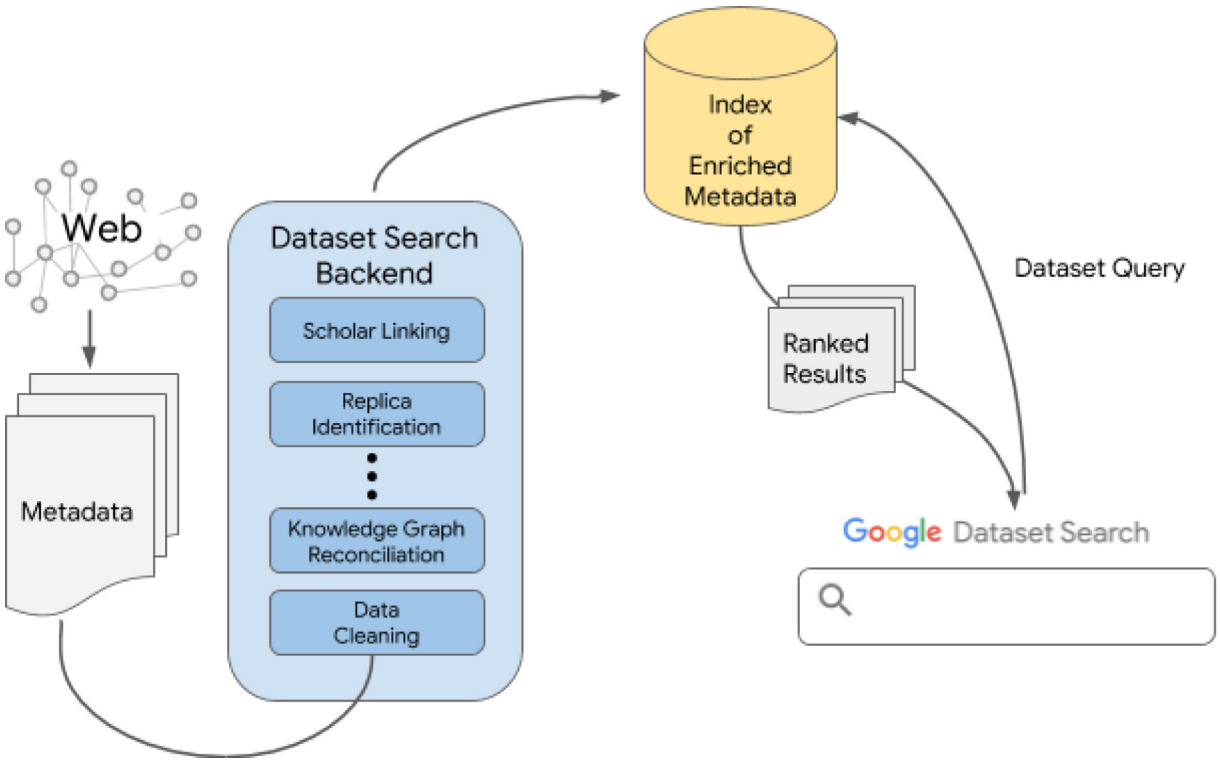


Fonte: (WIKIPEDIA, 2020).

Realizadas as buscas e levantamentos bibliográficos, prossegue-se com as definições e contextualizações. De acordo com Google (2019a), exemplos do que pode ser qualificado como um conjunto de dados representáveis por *metadados* são: uma tabela ou um arquivo CSV com alguns dados; um conjunto organizado de tabelas; um arquivo em formato proprietário que contenha dados; uma coleção de arquivos que unidos formam um conjunto de dados significativo; um objeto estruturado com dados em algum outro formato para processamento; imagens que capturam dados; arquivos relacionados ao aprendizado de máquina, como parâmetros treinados ou definições de estrutura de rede neural ou qualquer outro conjunto de dados representável e quantificável, em seus mais diversos formatos (HALEVY *et al.*, 2016): texto puro, planilhas, tabelas gigantes, sistemas de arquivos em nuvem, bases de dados relacionais etc. naturalmente ocasionando em uma ampla diversidade de metadados.

Colocadas estas considerações, o GOODS, segundo Noy (2020), surgiu partindo da dificuldade em encontrar repositórios de dados na internet. A autora afirma que, nos últimos anos, houve um aumento significativo em quantidade, volume e tamanho, além da proliferação e expansão de dados desestruturados na web, que afetou também o mundo científico e as ferramentas de busca, que não conseguem localizar dados no espectro chamado "cauda longa da pesquisa" disponíveis na internet. A proposta, portanto, foi a criação de um buscador, segundo (HALEVY *et al.*, 2016), que fosse capaz de coletar, organizar e indexar de acordo com os padrões vigentes metadados de *datasets* acessíveis pela internet. Com base nestas premissas, o GOODS foi concebido em meados de 2015, lançado em 2018 em versão beta e em 2020 como versão oficial. A figura 2 mostra uma visão geral da ferramenta.

Figura 2 – Visão geral do Google Dataset Search (GOODS)



Fonte: (BURGESS; NOY, 2018, on-line).

O protocolo "robots.txt", ou Protocolo de Exclusão de Robôs, é um conjunto de comandos que são interpretados por robôs de busca e indicam quais os diretórios e páginas de seu site não devem ser acessados por eles. Geralmente é disponibilizado na raiz das páginas web e contém comandos específicos sobre as URLs e conteúdos específicos do domínio. Tradicionalmente, por meio deste protocolo, 'robôs' digitais da *Google* coletam metadados de páginas web. Estes metadados são organizados, normalizados, indexados e organizados por prioridades para serem localizados por usuários através da interface de consulta (BURGESS; NOY, 2018). Graças aos padrões de metadados, a plataforma consegue identificar os *datasets*, conectar com outras ferramentas (como o *Google Scholar* e o *Google Knowledge Graph*) e assim extrair de forma otimizada a informação desejada. A indexação dos metadados permitem ainda eliminar *datasets* duplicados ou disponibilizados em lugares diferentes.

Com a saída da plataforma da versão BETA, novas funcionalidades foram incorporadas, além da possibilidade de democratizar o acesso a *datasets* para qualquer tipo de usuário (como por exemplo buscar "esqui" e encontrar *datasets* que abrangem desde a velocidades dos esquiadores mais rápidos às receitas dos resorts de esqui). Dentre as novas funcionalidades, estão a filtragem dos resultados de busca de acordo com o tipo de *dataset* que se deseja (ex.: tabelas, imagens, arquivos texto, se são gratuitos ou pagos etc), a possibilidade de acesso através de dispositivos móveis e melhorias de interface (NOY, 2020).

Segundo Canino (2019), é importante ressaltar a diferença entre o GOODS e outras ferramentas da *Google*. O primeiro provê uma busca mais profunda e detalhada em fontes distintas de dados, comparado com outras ferramentas da mesma empresa, como por exemplo o "*Google Public Data Explorer*", que suporta apenas formatos XML e CSV, ou do "*Google Knowledge Graph*", específico para visualização e conexão entre metadados. Ainda assim, a empresa afirma promover a interação entre os produtos com o intuito de aprimorar e testar algoritmos de busca já utilizados nas plataformas (NOY, 2020).

PADRÕES, DIRETRIZES E FORMATOS ACEITOS PARA INTEROPERABILIDADE ENTRE *DATASETS*

As orientações gerais e formatos para desenvolvedores, segundo *Google* (2019), dizem que é possível processar dados estruturados em páginas da Web sobre conjuntos de dados das seguintes formas: ou usando a marcação de conjunto de dados do *schema.org* ou estruturas equivalentes representadas no formato de vocabulário do catálogo de dados (DCAT, na sigla em inglês) do W3C (páginas em inglês). Também estão sendo testadas suportes experimentais para dados estruturados com base no CSVW do W3C. Para auxiliar a compreensão desta terminologia, o quadro 1 traz definições sobre as siglas utilizadas neste trabalho e encontradas nas diretrizes de interoperabilidade do GOODS. No entanto, a abordagem do *Google* para a descoberta de conjuntos de dados recomenda fortemente o uso das diretrizes da *schema.org*, que podem ser adicionados a páginas que descrevem conjuntos de dados.

Quadro 1 – Informações básicas sobre as siglas de formatos de dados utilizadas neste trabalho

SIGLA	DESCRIÇÃO
RDF	Resource Description Framework: representa meta dados no formato de sentenças sobre propriedades e relacionamentos entre itens na web.
JSON	Formato de interoperabilidade de dados entre sistemas, independente da linguagem de programação.
JSON-LD	JSON Linked data: é a maneira na qual a internet usa para conectar dados relacionados.
DCAT	Data catalog vocabular. Esquema de dados para facilitar a interoperabilidade entre dados de catálogos publicados na web.
URI	Uniform resource identifier: permite obter um identificador único para qualquer recurso na internet através de uma URL inteligente
Microdados	Conjunto de etiquetas de organização de conteúdos que são legíveis por computadores e pessoas.
CSVW	CSV on the web working group: utiliza o padrão RDF para dados tabulares

Fonte: Dados da pesquisa, 2019.

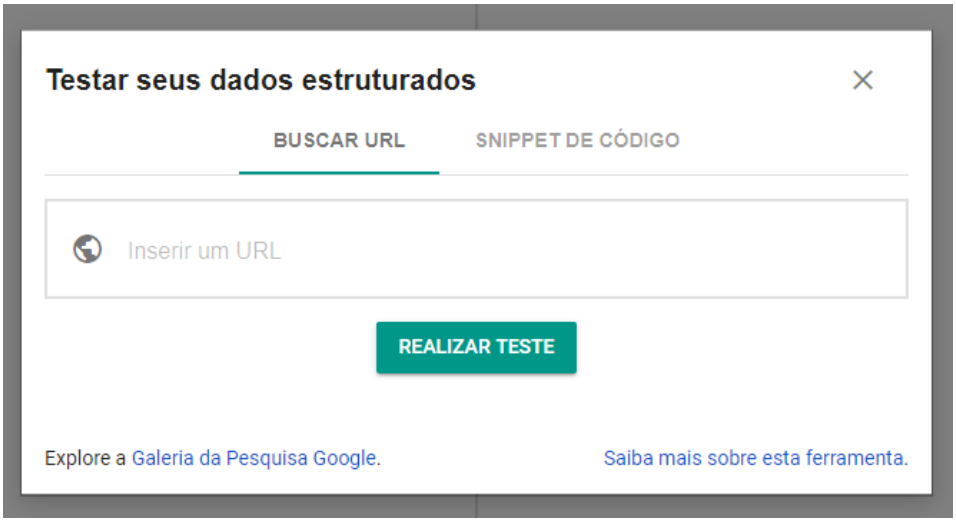
Criada pelo *Google*, Microsoft, Yahoo e Yandex, a *schema.org* é uma comunidade colaborativa com a missão de criar, manter e promover esquemas de dados estruturados na internet: em páginas web, mensagens de e-mail, conjuntos de dados e afins. O vocabulário *schema.org* pode ser utilizado em diversas codificações (como RDFa, Microdados e JSON-LD), e seu vocabulário compreende entidades, relacionamentos entre entidades e ações entre elas para estruturar e organizar dados. Segundo o site oficial, mais de 10 milhões de páginas utilizam *schema.org*, dentre elas aplicações da empresa *Google*. As orientações e vocabulários para estruturar conjuntos de dados do *schema.org* estão disponíveis em *https://schema.org/Dataset* e incluem variáveis como autor, data de publicação, palavras chave, link de acesso, codificação, licença, versão, linguagem etc.

O DCAT (*data catalog vocabulary*) é um vocabulário RDF criado pela W3C para facilitar a interoperabilidade entre catálogos de dados publicados na internet. O W3C (*World Wide Web Consortium*) atualmente é uma organização composta por 450 membros, dentre eles órgãos governamentais, empresas, organizações independentes e comunidades científicas, com a finalidade de estabelecer padrões para criação e interpretação de conteúdo na internet (W3C, 2019).

Assim, para serem inseridos no GOODS, os sites precisam seguir as diretrizes de dados estruturados (estar em JSON – padrão recomendado -, DCAT ou microdados). Além dessas diretrizes, o site oficial indica as práticas recomendadas de *sitemap* (arquivos que auxiliam o *Google* a encontrar as URLs do site) e origem e procedência – sobretudo quando conjuntos de dados abertos são republicados, agregados e baseados em outros conjuntos de dados.

Com o objetivo de facilitar a adesão aos padrões internacionais, a *Google* disponibiliza, no endereço eletrônico *https://search.google.com/structured-data/testing-tool*, uma ferramenta para realização de testes de dados estruturados, conforme ilustra a figura 3. A *Google* recomenda ainda a utilização de procedimentos como a Ferramenta de inspeção de URL (disponível no próprio console de busca fornecido pela empresa), para testar como o *Google* vê a página, o que acelera ainda mais o processo de indexação e disponibilização dos metadados nos resultados de busca do GOODS (caso existam *datasets* eleitos e disponíveis) e em outros serviços da *Google*.

Figura 3 – Ferramenta de teste de dados estruturados disponibilizada pela Google



Fonte: (GOOGLE, 2019b, on-line).

RASTREAMENTO E INSERÇÃO DE *DATASETS* NO *GOOGLE DATASET SEARCH*

O rastreamento dos sites que contém *datasets* é feito da mesma forma que as outras ferramentas da *Google*. Através do protocolo "robots.txt", os robôs buscadores ao acessarem o endereço do domínio podem ser orientados quanto ao controle de acesso e indexação de arquivos de imagem, a conteúdos em arquivos em geral (entram aqui os *datasets*), a arquivos de programação da própria página, o acesso aos mapas do site (sitemaps), dentre outros. Uma boa configuração do "robots.txt" e programação da página que contém o dataset conforme as diretrizes da schema.org é essencial para a indexação e disponibilização dos conjuntos de dados que se deseja inserir na plataforma de busca (GOOGLE, 2020).

TESTES: VISÃO GERAL DA FERRAMENTA

No tocante a utilização prática da ferramenta, o acesso pode ser feito pelo site https://datasetsearch.research.google.com. Ao acessar a página, percebe-se o mesmo padrão visual do buscador da empresa. O usuário conta com quatro opções distintas: 1) um campo de busca para realizar sua pesquisa por *datasets;* 2) um link "saiba mais" sobre como incluir conjuntos de dados nos resultados de busca; 3) um botão "sobre" com informações gerais acerca do GOODS e 4) um botão de feedback, para que os usuários possam informar eventuais problemas técnicos ou dificuldades de acesso.

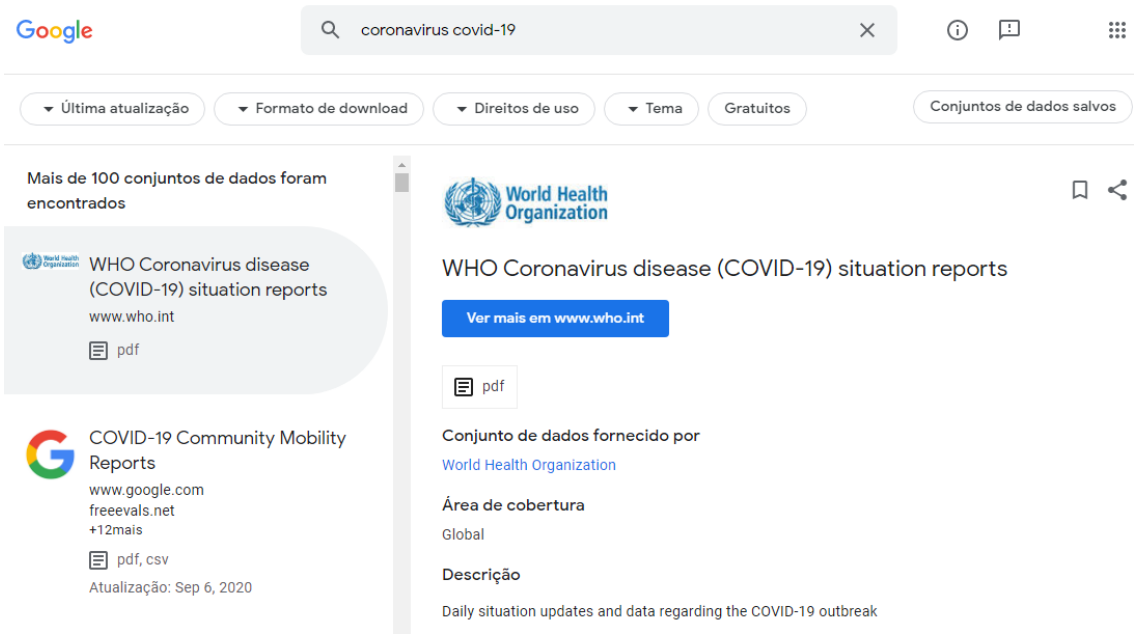
A página inicial permite ao usuário selecionar o idioma de utilização, tendo como padrão a linguagem definida pelo sistema operacional do usuário.

Os resultados de busca são dispostos da seguinte forma: à esquerda, com barra de rolagem, estão listados os *datasets* localizados conforme o critério de busca, sendo os mais relevantes primeiro e os menos relevantes em seguida. Ao clicar em um deles, são exibidos à direita da tela, nesta ordem (de cima para baixo): o título do *dataset*, o link para acesso aos dados, informações gerais como data de publicação e atualização, o fornecedor do conjunto de dados, a área de cobertura, licença, formato e descrição. A figura 4 ilustra a tela inicial do *Google Dataset Search* e a figura 5 mostra os resultados de busca. Ao visualizar os resultados de busca (figura 5), ficam visíveis uma das novidades da versão oficial (pós-beta), que são os filtros dos resultados de busca. São eles: por tempo de atualização/publicação, por formato de download, por direitos de uso, por tema e por tipo de acesso (gratuito ou não), além da possibilidade de salvar conjuntos de dados e vinculá-los à conta autenticada no momento para posterior análise.

Figura 4 – Tela inicial do *Google Dataset Search*

Fonte: (GOOGLE, 2019b, on-line).

Figura 5 – Resultados de busca pelo termo "coronavirus covid-19", sugerido pela página inicial



Fonte: Captura de tela da busca dos termos coronavírus covid-19 (GOOGLE, 2019b).

TESTES DE FUNCIONALIDADE

Para o teste de funcionalidade do tipo caixa-preta, e acatando a sugestão da própria ferramenta em sua página inicial, foi feita uma busca com o termo "*coronavirus covid-19*". Foram obtidos mais de 100 resultados indicando *datasets* contendo estes termos. No entanto, ao realizar outras buscas, percebemos que a ferramenta limita a quantidade de resultados a "mais de 100" não mostrando de fato o número real de itens obtidos. Um aspecto importante a ser mencionado é a eficiência e rapidez de indexação e disponibilização de *datasets*: diante da sugestão da própria ferramenta (covid-19) foi possível perceber que conjuntos de dados publicados em questão de poucos dias já estavam indexados e disponíveis para consulta.

Observou-se que o buscador também permite alguns comandos de busca avançada do buscador tradicional da *Google*, como operadores "site:" para restringir a busca a um domínio específico, ou "inurl:" para identificar um termo dentro de uma URL específica, os caracteres asterisco (*) para substituir por qualquer conteúdo e aspas duplas para encontrar frase exata.

Os cliques nas funções de visualização do conjunto de dados, do fornecedor e da visualização externa (no site de origem) dos 20 primeiros resultados de busca funcionaram adequadamente.

Dessa forma, os resultados obtidos no teste caixa-preta demonstraram que a ferramenta, de acordo com suas funções, entrega o que propõe, sendo capaz de buscar, listar e viabilizar conjuntos de dados já indexados pela plataforma.

TESTES DE INTERFACE E USABILIDADE

Observou-se que o padrão de interface e interação com o usuário segue o padrão Material Design, linguagem de design e padrão visual desenvolvido em 2014 pela *Google*, utilizando layouts baseados em grids, animações e transições responsivas, preenchimentos, e efeitos de profundidade como luzes e sombras, nas cores predominantemente azul, vermelho e branco.

O quadro a seguir mostra o resultado da aplicação, a título de visão geral, das 10 heurísticas de Nielsen (2014) no *Google Dataset Search*, considerando a letra C para alto grau de violação (heurística comprometida), a letra B para violação parcial e A nenhuma violação (heurística preservada):

Quadro 2 − Aplicação das 10 heurísticas propostas por Nielsen aplicadas ao GOODS

HEURÍSTICA	GRAU	OBSERVAÇÃO
Visibilidade do status do sistema	B	Não mostra o total exato de datasets obtidos
Compatibilidade entre o sistema e o mundo real	A	Familiaridade com buscador Google
Controle e liberdade para o usuário	A	Permite abrir links em novas janelas, cancelar buscas e baixar diretamente os arquivos etc.
Consistência e padronização	A	Segue os mesmos padrões visuais de todos os produtos da empresa.
Prevenção de erros	A	Não foram encontrados erros durante os testes.
Reconhecimento ao invés de memorização	A	Interface simples e intuitiva.
Eficiência e flexibilidade de uso	B	Os testes demonstraram rapidez e agilidade para retornar os resultados. Porém, não há campos de busca avançada ou instruções explícitas para utilização de operadores de busca.
Estética e Design minimalista	A	Apresenta conteúdo relevante e funcional para o contexto.
Auxilia os usuários a reconhecerem, diagnosticarem e recuperarem-se de erros	A	Não foram encontrados erros, mas existe botão de feedback nas telas.
Ajuda e documentação	A	A ferramenta disponibiliza links de ajuda e instruções gerais de uso.

Fonte: Dados da pesquisa, 2019.

A aplicação das heurísticas obteve 80% de não violação, apontando preliminarmente que a interface e usabilidade do *Google Dataset Search* atende de maneira satisfatória aos seus usuários. O resultado não poderia ser diferente, uma vez que o modelo mental proporcionado pelo padrão *material design* já é bastante difundido e reconhecido por grande parte das pessoas que utilizam as ferramentas *Google*, incluindo o sistema operacional *Android*, presente em 85,4% dos smartphones em operação até a presente data (IDC, 2020).

TESTES DE BUSCA EM DOMÍNIOS E ÁREAS ESPECÍFICAS

De acordo com Noy (2020) as áreas que possuem mais volumes de *datasets* são: geociências, biologia e agricultura. A área governamental também merece destaque, visto que grande parte dos governos publicam dados de acordo com os padrões estabelecidos pela schema.org. Mais de 2 milhões de *datasets* governamentais norte-americanos estão disponíveis. A pesquisadora ainda ressalta que o tipo de *datasets* mais presentes na plataforma são tabelas. As figuras 6, 7 e 8 trazem mais informações estatísticas sobre o GOODS.

Uma vez mencionados dados governamentais e, para fins de testes exploratórios, foram realizadas, em meados de setembro de 2020, consultas de testes em sites e domínios governamentais do Brasil para se ter uma noção geral da quantidade de conjuntos de dados mapeados no espaço eletrônico brasileiro. Os termos utilizados, a quantidade de resultados encontrados e os itens de maior relevância (segundo ranking do GOODS) foram listados no quadro 3.

Figura 6 – Estatísticas do GOODS. Em a, domínios com maior número de datasets, responsáveis por 65% dos conjuntos de dados disponíveis. Em b, quantitativos de datasets por domínios e em c quantidade de datasets por idiomas

a)

Domain	Datasets
ceicdata.com	3.7M
data.gov	3.1M
hikersbay.com	2.3M
tradingeconomics.com	2.2M
knoema.com	1.7M
figshare.com	1.3M
stlouisfed.org	1.2M
datacite.org	1.1M
thermofisher.com	1.0M
statista.com	0.9M

b)

Top-level domain	Number of datasets
.com	14,956K
.org	4,696K
.gov	3.386K
.at	819K
.net	760K
.es	524K
.de	366K
.edu	293K
.fr	281K
.eu	263K

c)

Language	Number of datasets	% increase
English	18,650K	67%
Chinese	1,851K	82%
Spanish	1,485K	70%
German	743K	74%
French	492K	76%
Arabic	435K	75%
Japanese	404K	72%
Russian	354K	65%
Portuguese	304K	69%
Hindi	288K	70%

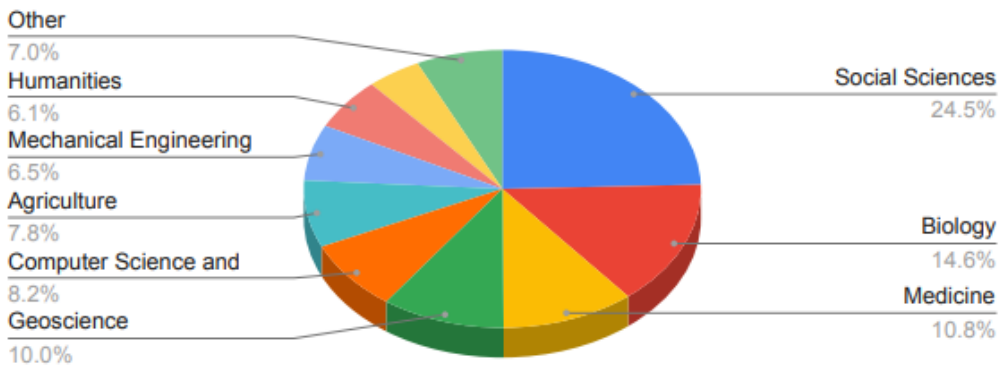
Fonte: (BENJELLOUN; CHEN; NOY, 2020).

Figura 7 – Percentuais de conjuntos de datasets agrupados por formatos

Category	Number of datasets	% of total	Sample formats
Tables	7,822K	37%	CSV, XLS
Structured	6,312K	30%	JSON, XML, OWL, RDF
Documents	2,277K	11%	PDF, DOC, HTML
Images	1,027K	5%	JPEG, PNG, TIFF
Archives	659K	3%	ZIP, TAR, RAR
Text	623K	3%	TXT, ASCII
Geospatial	376K	2%	SHP, GEOJSON, KML
Computational biology	110K	<1%	SBML, BIOPAX2, SBGN
Audio	27K	<1%	WAV, MP3, OGG
Video	9K	<1%	AVI, MPG
Presentations	7K	<1%	PPTX
Medical imaging	4K	<1%	NII, DCM
Other categories	2,245K	11%	

Fonte: (BENJELLOUN; CHEN; NOY, 2020).

Figura 8 – Percentuais de datasets organizados por áreas de conhecimento



Fonte: (BENJELLOUN; CHEN; NOY, 2020).

Quadro 3 − Amostra de testes em sites de domínio.br realizados no GOODS

	TERMO DE BUSCA	RESULTADOS	ITENS MAIS RELEVANTES
A	site:*.gov.br	> 100	ana.gov.br e hub.arcgis.com (dados ambientais, geografia e agricultura)
B	inurl:*gov.br	> 100	data.wu.ac.at (indicadores de gestão e infraestrutura)
C	inurl:*edu.br	> 100	cloud.csiss.gmu.edu (informações ligadas a graduação e pós-graduação)
D	inurl:*com.br	> 100	dados comerciais
E	site:"*.org.br"	57	Dados de eventos e monitoramentos em geral
F	universidade	> 100	Dados biológicos - GBIF
G	"são paulo"	> 100	Diversos (dados governamentais e repositórios em universidades).

Fonte: Termos de pesquisados em 2020 no buscador Google Dataset Search (GOOGLE, 2019b).

Após análise, os resultados de busca em (A e B) apontam que grande parte dos dados indexados pelo GOODS são provenientes de portal de dados abertos[1] da Agência Nacional de Águas (ANA), que oferece ferramentas para utilização de dados públicos sobre recursos hídricos no Brasil. Dados geográficos, mapeamentos e afins também estão disponíveis em hub.arcgis.com. Outro fato interessante é que muitos conjuntos de dados já estão indexados e disponíveis na plataforma csiss.gmu.edu, que pertence à GEOSS *Information Exchange DataHub*. Tal ferramenta informa que existem, na data deste trabalho, 15.223 conjuntos de dados etiquetados com "BRASIL" e 7.852 com a etiqueta "IBGE", mostrando forte relevância para a indexação de conjuntos de dados no Brasil, especialmente em itens relacionados a gestão e infraestrutura. Em C nota-se que a ferramenta indexou majoritariamente dados de instituições de ensino superior ligados à graduação e pós-graduação. Observou-se que em D o GOODS indexou tabelas de preços, listas de clientes e até especificações de produtos disponíveis em e-commerce. Em E foram localizados dados ligados a organização de cursos e eventos, mapas e resultados de monitoramentos para fins específicos (geográficos e biológicos).

Em F, observou-se que existem diversas universidades brasileiras que disponibilizam seus dados pela plataforma internacional GBIF - Sistema Global de Informação sobre Biodiversidade. Em G tentou-se identificar *datasets* relacionados ao Estado mais populoso do Brasil. Foram encontrados repositórios governamentais ligados a ANA, Embrapa, em plataformas como a *Zenodo, Kaggle* e *ReSearchGate*, dados comerciais isolados e alguns repositórios de universidades públicas.

Merece destaque o fato de que os testes realizados confirmam as publicações oficiais da *Google*, que apontam a maioridade em volume de dados de cunho social, nas áreas de geociências e biologia, com destaque para dados governamentais.

Com relação ao formato dos *datasets* encontrados, os que mais se sobressaíram nos resultados de busca expostos pelos termos do quadro 2 foram *xls, csv e json*, além de API's próprias para consulta e exploração dos dados, confirmando as estatísticas descritas por Benjelloun, Chen e Noy (2020).

É interessante notar também que já se encontram indexados e disponíveis vários *datasets* provenientes do Portal Brasileiro de Dados Abertos (dados.gov.br), uma vez que seguem os mesmos padrões e formatos seguidos pelo GOODS.

[1] Disponível em: dadosabertos.ana.gov.br

Por fim, outro fator relevante é que, naturalmente, os itens catalogados em português ainda são poucos (BENJELLOUN; CHEN; NOY, 2020) se comparados ao idioma inglês, embora muitos conjuntos de dados brasileiros estejam de acordo com os padrões exigidos pelo GOODS para indexação e disponibilização.

CONSIDERAÇÕES FINAIS

Os levantamentos bibliográficos apontam que as ferramentas *Google* desempenham importante papel como agentes das transformações informacionais que a sociedade de hoje vivencia. Desde o tradicional buscador às ferramentas voltadas para a comunidade científica, a *Google* vem, entre prós e contras (CANINO, 2019), galgando espaços significativos como instrumento de acesso e difusão da informação.

A proposta da *Google* para indexação e organização de *datasets* por meio do GOODS apresenta-se como promissora, considerando o histórico de ferramentas lançadas pela referida empresa e sua participação no mercado de tecnologia, padrões internacionais de dados e a notável capacidade de seu aparato computacional e tecnologias, como infraestrutura em nuvem, mineração de dados e inteligência artificial.

Partindo desta premissa, acredita-se num não tardio aumento de integrações com bases nacionais de conjuntos de dados, universidades e indústrias, sobretudo as que não possuem ferramentas para armazenar e distribuir seus *datasets*. Esta será uma grande contribuição para expor parte da cauda longa de pesquisa, tão importante para a comunidade científica. Quanto aos desafios e oportunidades, Goben e Sandusky, (2020) apontam que ainda existem diversos entraves técnicos, financeiros einformacionais no provimento de acesso a dados abertos, sendo para isso fundamental a participação de profissionais como bibliotecários, cientistas da informação, analistas e políticos, bem como a disponibilização de recursos financeiros por parte de instituições credenciadas, a fim de fazer com que os *datasets* – importantes ferramentas para a comunidade científica - possam ser compartilhados de maneira apropriada e assim possam chegar aos respectivos interessados.

Especificamente quanto ao GOODS, Canino (2019) conjectura que o futuro da ferramenta ainda é impreciso, considerando ou uma ascensão meteórica, ou apenas mais um dos muitos produtos *Google* que não vingaram, não passando nem mesmo da fase BETA. No entanto, podemos perceber que o estágio atual do GOODS já não é mais beta, indicando pela quantidade ascendente de *datasets* catalogados que a ferramenta está em pleno crescimento. Canino (2019) ressalta que uma interface simples de busca, interoperável com os principais padrões de dados e suportado pela *Google* tem total potencial para tornar a descoberta de relevantes conjuntos de dados uma tarefa simples, rápida e eficiente, transformando assim de modo positivo o jeito de lidar com dados brutos em pesquisas científicas e tornando-se referência nesta tarefa.

Por fim, este breve estudo, ao oferecer uma visão geral e perspectivas para indexação e disponibilização de conjuntos de dados científicos abertos através da plataforma *Google Dataset Search*, cumpre os objetivos a que foi proposto.

REFERÊNCIAS

BENJELLOUN, O.; CHEN, S.; NOY, N. Google Dataset Search by the Numbers. *arXiv:2006.06894 [cs]*, 11 jun. 2020. Disponível em: http://arxiv.org/abs/2006.06894. Acesso em: 10 mar. 2021.

BRASIL. *Portal Brasileiro de Dados Abertos.* 2019. Disponível em: http://dados.gov.br. Acesso em: 13 set. 2019.

BURGESS, M.; NOY, N. *Building Google Dataset Search and Fostering an Open Data Ecosystem*. 26 set. 2018. *Google AI Blog: The latest news from Google AI.* Disponível em: https://ai.googleblog.com/2018/09/building-google-dataset-search-and.html. Acesso em: 10 set. 2019.

CANINO, A. Deconstructing Google Dataset Search. *Public Services Quarterly*, v. 15, n. 3, p. 248–255, 3 jul. 2019. DOI 10.1080/15228959.2019.1621793. Disponível em: https://www.tandfonline.com/doi/full/10.1080/15228959.2019.1621793. Acesso em: 10 mar. 2021.

GAVRON, E. M.; CANTO, F. L. do. Análise da utilização dos periódicos de acesso aberto de uma base de dados assinada pela biblioteca universitária da UFSC. *In*: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA, DOCUMENTAÇÃO E CIÊNCIA DA INFORMAÇÃO (CBBD), 27., 2017. *Anais* [...]. Fortaleza: FEBAB, 2017. v. 27, p. 1–6. Disponível em: https://portal.febab.org.br/anais/article/view/1787. Acesso em: 13 set. 2019.

GERHARDT, T. E.; SILVEIRA, D. T. (Orgs.). *Métodos de pesquisa.* Porto Alegre: Editora da UFRGS, 2009. Disponível em: http://www.ufrgs.br/cursopgdr/downloadsSerie/derad005.pdf. Acesso em: 12 set. 2019.

GOBEN, A.; SANDUSKY, R. J. Open data repositories: Current risks and opportunities. *College & Research Libraries News*, v. 81, n. 2, p. 62, 4 fev. 2020. DOI 10.5860/crln.81.1.62. Disponível em: https://crln.acrl.org/index.php/crlnews/article/view/24273. Acesso em: 10 mar. 2021.

GOOGLE. *Conjuntos de diretrizes e orientações sobre o Google Dataset Search.* 2019a. *Google Search Central*. Disponível em: https://developers.google.com/search/docs/data-types/dataset?hl=pt-br. Acesso em: 13 set. 2019.

GOOGLE. *Dataset Search.* 2019b. *Google*. Disponível em: https://datasetsearch.research.google.com/. Acesso em: 12 set. 2019.

GOOGLE. *Rastreamento e indexação: manual de orientações técnicas para criação de metadados para rastreio de páginas web*. 2020. *Google Search Central.* Disponível em: https://developers.Google.com/Search/reference/robots_meta_tag. Acesso em: 20 abr. 2020.

GOOGLE. *Testar seus dados estruturados.* 2019c. **G***oogle: Ferramenta de teste de dados estruturados.* Disponível em: https://search.google.com/structured-data/testing-tool. Acesso em: 12 set. 2019.

HALEVY, A.; KORN, F.; NOY, N. F.; OLSTON, C.; POLYZOTIS, N.; ROY, S.; WHANG, S. E. Goods: Organizing Google's Datasets. *In*: SIGMOD/PODS'16: INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 14 jun. 2016. *Proceedings of the 2016 International Conference on Management of Data* [...]. San Francisco California USA: ACM, 14 jun. 2016. p. 795–806. DOI 10.1145/2882903.2903730. Disponível em: https://dl.acm.org/doi/10.1145/2882903.2903730. Acesso em: 10 mar. 2021.

INTERNATIONAL DATA CORPORATION (IDC). *Smartphone Market Share.* 15 dez. 2020. *IDC*. Disponível em: https://www.idc.com/promo/smartphone-market-share/os. Acesso em: 2 set. 2020.

MYERS, G. J.; SANDLER, C.; BADGETT, T. *The art of software testing*. 3rd ed. Hoboken, N.J: John Wiley & Sons, 2012.

NIELSEN, M. A. *Reinventing discovery: the new era of networked science.* [*S. l.*: *s. n.*], 2014. Disponível em: http://www.vlebooks.com/vleweb/product/openreader?id=none&isbn=9781400839452. Acesso em: 9 set. 2020.

NOY, N. *Discovering millions of datasets on the web.* 23 jan. 2020. *Google: The keyword.* Disponível em: https://blog.google/products/search/discovering-millions-datasets-web/. Acesso em: 20 abr. 2020.

ROSA, J. M.; VERAS, M. Avaliação heurística de usabilidade em jornais on-line: estudo de caso em dois sites. *Perspectivas em Ciência da Informação*, v. 18, n. 1, p. 138–157, mar. 2013. DOI 10.1590/S1413-99362013000100010. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362013000100010&lng=pt&tlng=pt. Acesso em: 10 mar. 2021.

WIKIPEDIA. *Visualizações da página: comparação das visualizações entre várias páginas.* 10 set. 2020. *Visualizações.* Disponível em: https://en.wikipedia.org/wiki/Google_Dataset_Search. Acesso em: 10 set. 2020.

WORLD WIDE WEB CONSORTIUM (W3C). *Current Members.* 2019. *World Wide Web Consortium.* Disponível em: https://www.w3.org/Consortium/Member/List. Acesso em: 9 set. 2019.

W3C - World Wide Web Consortium. *Data Catalog Vocabulary* (DCAT). 2014. Disponível em: https://www.w3.org/TR/vocab-dcat/. Acesso em: 09 set. 2019.

Perfil das orientações e produções das mulheres fundamentado em dados da Plataforma Lattes

Monique de Oliveira Santiago
Mestranda em Modelagem Matemática e Computacional pelo Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) - Belo Horizonte, MG - Brasil. Especialização em Análise de Dados com BI e Big Data pela Universidade Cruzeiro do Sul (UNICSUL) - Brasil. Professora da Universidade do Estado de Minas Gerais (UEMG) - Brasil.
http://lattes.cnpq.br/3530976051984613
E-mail: moniqueosantiago@gmail.com

Felipe Affonso
Mestrando em Modelagem Matemática e Computacional pelo Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) - Belo Horizonte, MG - Brasil. Graduação em Engenharia da Computação pela Universidade do Estado de Minas Gerais (UEMG) – Brasil, com período sanduíche em University of Miami (Umiami) - Estados Unidos.
http://lattes.cnpq.br/1468618041970656
E-mail: felipe-affonso@hotmail.com

Thiago Magela Rodrigues Dias
Doutor em Modelagem Matemática e Computacional pela Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Belo Horizonte, Minas Gerais, Brasil. Professor do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) - Belo Horizonte, MG - Brasil.
http://lattes.cnpq.br/4687858846001290
E-mail: thiagomagela@gmail.com

Submetido em:25/09/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021 .

RESUMO

A temática mulheres na ciência tem sido foco de diversos estudos que buscam entender o papel da mulher e suas variadas relações com a ciência, trazendo abordagens que analisam a participação científica e tecnológica e buscam compreender sua trajetória acadêmica na ciência. Neste contexto, este trabalho objetivou analisar a participação das mulheres, utilizando como base de dados o conjunto de doutoras que possuem currículos cadastrados na Plataforma Lattes e cujo gênero registrado em seu currículo seja do sexo feminino. Após a coleta dos dados, obteve-se um conjunto de 125.515 currículos cadastrados com gênero feminino e titulação máxima concluída (doutorado) distribuída nas diversas áreas do conhecimento científico. Os dados das doutoras foram agrupados quanto a orientações e a produções bibliográficas e técnicas, nos quais foi possível analisar a evolução da produção científica e tecnológica do conjunto de forma temporal. Estudar os diversos aspectos da mulher em geral e particularmente na ciência, além de ser relevante, apresenta uma caracterização de suas pesquisas, podendo contribuir para gerar indicadores científicos nacionais e para a gestão das informações na área científica e tecnológica, bem como, para formulação de políticas e estratégias que potencializem a atuação das mulheres nos ambientes acadêmicos.

Palavras-chave: Mulheres na ciência. Bibliometria. Produção científica e tecnológica. Gênero feminino.

Profile of women's guidelines and productions based on data from the Lattes Platform

ABSTRACT

The theme of women in science has been the focus of several studies that seek to understand the role of women and their varied relations with science, bringing approaches that analyze scientific and technological participation and seek to understand their academic trajectory in science. In this context, this paper aimed to analyze the participation of women using as a database the set of PhDs who have curricula registered in the Lattes Platform and whose gender registered in their curriculum is female. After completing the stage of data collection, a set of 125,515 registered curricula with female gender and maximum completed doctorate degree was obtained, distributed in the various areas of scientific knowledge. The data of the female PhDs were grouped as to orientations and bibliographic and technical productions, in which it was possible to analyze the evolution of the scientific and technological production of the set in a temporal way. Studying the various aspects of women in general and particularly in science, in addition to being relevant, presents a characterization of their research, which can contribute to the generation of national scientific indicators and the management of information in the scientific and technological area, as well as supporting the formulation of policies and strategies that enhance the performance of women in academic environments.

Keywords: *Women in science. Bibliometry. Scientific and technological production. Feminine gender.*

Perfil de las pautas y producciones de mujeres basado en datos de la Plataforma Lattes

RESUMEN

El tema mujeres en la ciencia ha sido el foco de varios estudios que buscan comprender el rol de la mujer y sus variadas relaciones con la ciencia, aportando enfoques que analizan la participación científica y tecnológica y buscan comprender su trayectoria académica en la ciencia. En este contexto, este trabajo tuvo como objetivo analizar la participación de las mujeres utilizando como base de datos el conjunto de médicos que tienen currículos registrados en la Plataforma Lattes y cuyo género registrado en su currículo es femenino. Tras la recogida de datos se obtuvo un conjunto de 125.515 planes de estudio registrados con género femenino y grado máximo de doctorado cursado, distribuidos en las distintas áreas del conocimiento científico. Los datos de los médicos se agruparon según pautas y producciones bibliográficas y técnicas, en las que se pudo analizar la evolución de la producción científica y tecnológica del conjunto de forma temporal. El estudio de los diversos aspectos de la mujer en general y particularmente en la ciencia, además de ser relevante, presenta una caracterización de su investigación, que puede contribuir a la generación de indicadores científicos nacionales y al manejo de la información en el área científica y tecnológica, así como para la formulación de políticas y estrategias que mejoren el desempeño de las mujeres en entornos académicos.

Palabras clave: *Mujeres en la ciencia. Bibliometría. Producción científica y tecnológica. Género femenino.*

INTRODUÇÃO

A ciência pode ser considerada uma ferramenta essencial para a busca de conhecimento, que objetiva compreender, questionar e aprimorar os processos e métodos científicos, assim como estudar os fenômenos da natureza e suas relações e também responder às necessidades da sociedade (LOPES *et al.*, 2020). A produção científica é resultado desta investigação científica e parte integrante do processo de conhecimento. O desenvolvimento de diversos estudos proporcionou um considerável crescimento das produções nos últimos anos, no qual dentre todos, a temática que compreende como a ciência tem evoluído e como a colaboração científica ocorre teve considerável destaque no meio científico. Normalmente, os dados relacionados à produção científica estão presentes em diversos repositórios, dificultando, assim, a recuperação e a análise dos dados. Entretanto, em um contexto brasileiro, esse processo pode ser facilitado pela utilização do repositório de dados curriculares da Plataforma Lattes, que é considerada um importante conjunto de dados científicos nacional, fornecendo informações de alta qualidade e possibilitando pesquisar dados dos indivíduos que estão ali cadastrados, como formação acadêmica e produção científica, dentre outros (LANE, 2010). Esse conjunto de dados integra, em um único sistema, as bases de dados de Currículos, de Grupos de Pesquisa e de Instituições de Ensino do país, com destaque para as informações curriculares, contendo a trajetória acadêmica de grande parte da comunidade científica brasileira e, por isso, pode ser utilizado para compreender a evolução da ciência, a produção e a colaboração científica.

Em geral, os grandes repositórios de dados disponíveis atualmente possuem características próprias, padrões únicos, quantidade e diversidade de dados, ocasionando, assim, uma tarefa complexa para realizar um estudo e explorar esses dados. Neste cenário, tem sido utilizada a bibliometria, que auxilia no processo de quantificação da comunicação escrita, utilizando métodos para análises estatísticas sobre a produção e a disseminação do conhecimento aplicado a fontes de dados científicos (ARAÚJO, 2006) e permite dispor indicadores para o planejamento nacional e a evolução das pesquisas científicas.

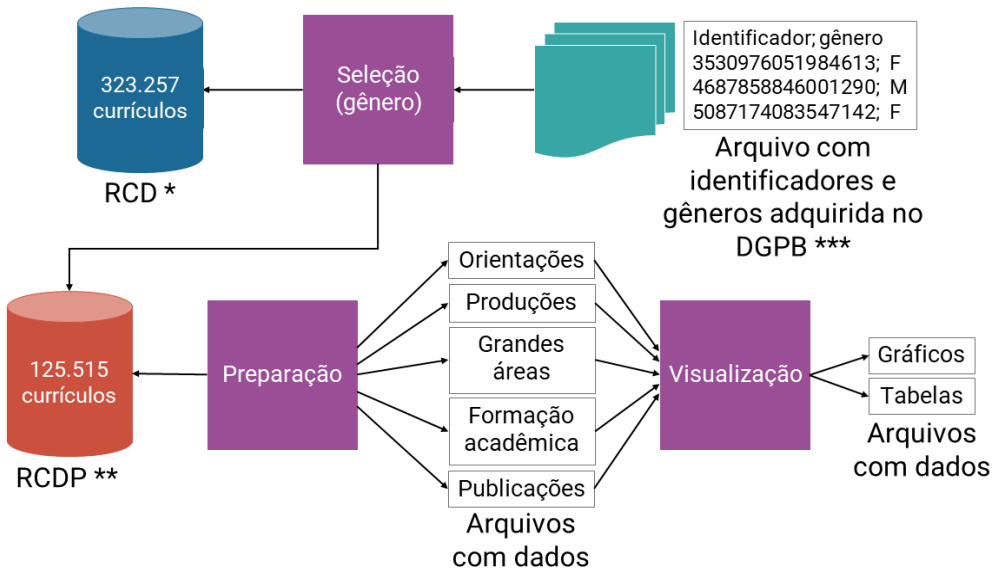
Dentre esses diversos estudos relacionados à produção científica, um que tem recebido destaque entre os pesquisadores refere-se ao gênero. Segundo Leta (2014), os estudos de gênero correspondem a um campo interdisciplinar que possui como temática a identidade e a representação de homens e mulheres na sociedade. Este campo inclui o subcampo Estudo da Mulher que compreende, dentre diversos aspectos, a mulher e suas variadas relações com a ciência. A temática refere-se à crescente participação das mulheres nas carreiras científica e tecnológica, adquirindo destaque entre os pesquisadores que buscam traçar um perfil da trajetória e do desempenho da mulher na ciência.

Realizar um estudo para compreender a participação científica das mulheres é uma iniciativa significativa para entender as desigualdades existentes, pois apesar do progresso na participação das mulheres em vários segmentos na carreira acadêmica e científica, ainda se percebe uma lacuna de gênero na ciência que precisa ser melhor compreendida. Assim, esse trabalho objetiva analisar a participação científica e tecnológica das doutoras pesquisadoras brasileiras, investigando como suas pesquisas têm sido realizadas e como sua trajetória acadêmica tem evoluído ao longo dos anos a partir de análises bibliométricas realizadas sobre dados curriculares disponíveis na Plataforma Lattes. Esse estudo, além de apresentar uma visão das mulheres que têm realizado pesquisas, visa a apresentar um cenário geral que pode contribuir para gerar indicadores científicos nacionais e para a gestão das informações na área científica e tecnológica.

METODOLOGIA

A Plataforma Lattes é um repositório de acesso aberto, que contém as informações curriculares incluídas pelo próprio indivíduo e estão disponíveis através da Web. Para analisar este amplo conjunto de dados, passou-se inicialmente pela etapa de aquisição dos dados e seleção dos currículos pelo critério de formação acadêmica/titulação e, em seguida, foi realizada a seleção pelo critério de gênero. Posteriormente, foi realizada a preparação dos dados a fim de gerar os arquivos específicos e, por fim, obter visualizações desses dados (figura 1).

Figura 1 – Processo de seleção pelo critério de gênero, preparação e visualização dos dados



Fonte: Elaboração dos autores.

Notas: * Repositório de Currículo de todos Doutores,
** Repositório de Currículos das Doutoras Pesquisadoras,
*** Diretório dos Grupos de Pesquisa no Brasil

Assim, a primeira etapa que corresponde ao processo de aquisição dos dados e seleção dos currículos pelo critério de formação acadêmica/titulação, foi realizada através do arcabouço *LattesDataXplorer* desenvolvido por Dias (2016). Esse extrator possui um conjunto de técnicas e métodos responsáveis por coletar, selecionar, tratar e analisar os dados curriculares da Plataforma Lattes. A coleta de todos os currículos utilizando o arcabouço foi realizada em outubro de 2019, ultrapassando 6.300.000 registros. Os currículos são armazenados localmente e possuem o formato XML (*eXtensible Markup Language* - uma linguagem de marcação que contém seções e campos bem delimitados), sendo este formato de arquivo mais adequado para o processamento automático dos dados. Para realizar uma análise detalhada da participação científica das mulheres, optou-se por limitar os dados através do nível de formação acadêmica/titulação, reduzindo o conjunto para indivíduos que possuem o nível de formação doutorado concluído.

Apesar deste conjunto não ser o mais significativo entre os níveis de formação, conforme enunciado por Dias (2016), eles são responsáveis por 74,51% dos artigos publicados em periódicos e 64,67% dos artigos publicados em anais de congresso, além de possuir, em geral, data de atualização de seus currículos recente e, notadamente, são responsáveis pelo mais alto nível de formação, a saber, mestrado e doutorado.Logo, após a aquisição de todos os currículos cadastrados na Plataforma Lattes, foi utilizado o módulo de seleção do *LattesDataXplorer* para selecionar, dentre estes, os currículos que possuem a formação acadêmica/titulação doutorado concluído, totalizando assim um conjunto com 323.257 currículos armazenados no Repositório de Currículos de todos Doutores (RCD).

Identificar o sexo dos autores é um grande desafio quando abordamos a produção científica e os estudos métricos (LETA, 2014), e, como a Plataforma Lattes não disponibiliza o campo sexo para a consulta pública, para realizar a segunda etapa que corresponde à seleção pelo critério de gênero, foi necessário buscar essa informação em outra base de dados.

Assim, foi utilizado o Diretório dos Grupos de Pesquisa no Brasil (DGPB), que possui o inventário dos grupos de pesquisa científica e tecnológica em atividade no país, disponibilizando os dados dos Censos em formato aberto (DIRETÓRIO DO GRUPO DE PESQUISA NO BRASIL, 2019). Os arquivos foram processados, por estas vias, obtiveram-se os dados de identificador e gênero para todos os pesquisadores, estudantes e técnicos dos arquivos dos Censos. Portanto, uma lista de identificadores do gênero feminino foi obtida, através dela foi possível buscar no RCD os que possuem identificadores iguais aos da lista, definindo, assim, o conjunto principal de arquivos XML com total de 125.515 currículos, denominado Repositório de Currículos das Doutoras Pesquisadoras (RCDP).

Para a terceira etapa, referente à preparação dos dados, inicialmente foi identificada a estrutura dos arquivos do RCDP. Um arquivo XML é organizado através dos elementos e atributos, iniciando pelo elemento raiz (pai) e percorrendo os elementos filhos até alcançar os atributos. A estrutura dos currículos XML da Plataforma Lattes possui o elemento raiz, nomeado por "Curriculo-vitae", e inúmeros elementos filhos que possuem seus próprios elementos e atributos. Os arquivos XML extraídos da Plataforma Lattes apresentam ampla diversidade dos dados como, nível de formação acadêmica/titulação, grandes áreas de atuação, projetos de pesquisa e extensão, idiomas, produções bibliográficas e técnicas como artigos publicados em anais de congresso e periódico, apresentação de trabalhos, participação em bancas, eventos, orientações, dentre outros. Assim, acessando cada currículo XML e obtendo a informação específica daquele currículo, foi armazenado em coleções de arquivos estruturados. Como cada currículo possui uma quantidade específica de informações, esses dados foram agrupados quanto a orientações, produções, grandes áreas, formação acadêmica, publicações. Após agrupar os dados, foi realizada a última etapa que corresponde a visualizações dos dados, ou seja, à caracterização para facilitar as análises dos dados gerando gráficos e tabelas.

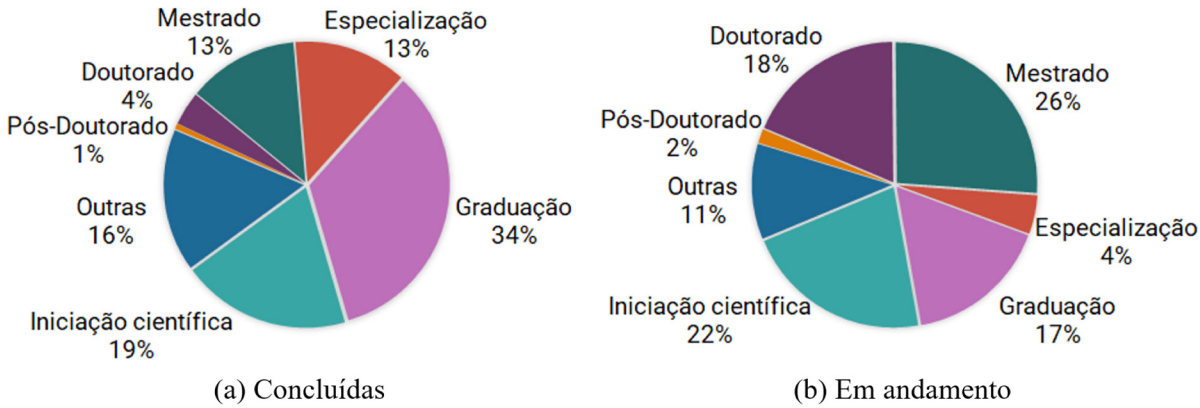
Quanto à limitação, no que se refere à coleta dos dados, deve-se considerar a confiabilidade dos dados, porém dois fatores impossibilitaram utilizar todo o conjunto de currículo das doutoras para análise. O primeiro corresponde à Plataforma Lattes que não disponibiliza as informações quanto ao gênero e o segundo corresponde ao DGPB, pois não são todas as doutoras que estão inseridas em um grupo de pesquisa. Vale ressaltar que a amostra utilizada neste estudo corresponde a todas as doutoras que estão inseridas ou participaram em algum momento de um grupo de pesquisa disponibilizado nos arquivos censitários do DGPB.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

Logo, os dados coletados da Plataforma Lattes, utilizando o arcabouço *LattesDataXplorer* em outubro de 2019, totalizaram mais de 6.300.000 registros. Desse total, foram selecionados os registros com o nível de formação acadêmica/titulação doutorado concluído, totalizando em 323.257 (5,13%) currículos de todos os doutores das diversas áreas do conhecimento científico. Esses mesmos dados foram selecionados pelo critério de gênero (utilizando o DGPB como validação), totalizando 125.515 (1,99%) currículos das doutoras pesquisadoras (RCDP), sendo que esses dados serão utilizados para as análises que retratam a participação científica feminina.

Como os doutores são responsáveis pela formação dos alunos em diferentes níveis de escolaridade, uma informação relevante para análise corresponde às orientações concluídas e em andamento. As orientações concluídas correspondem a todas as orientações realizadas pelas doutoras desde o início de sua carreira e que já estão finalizadas, totalizando 3.455.229 (figura 2 (a)). Enquanto as orientações em andamento são aquelas que estão em desenvolvimento e ainda não foram finalizadas, totalizando 334.570 (figura 2 (b)).

Figura 2 – Orientações por categorias



Fonte: Elaboração dos autores.

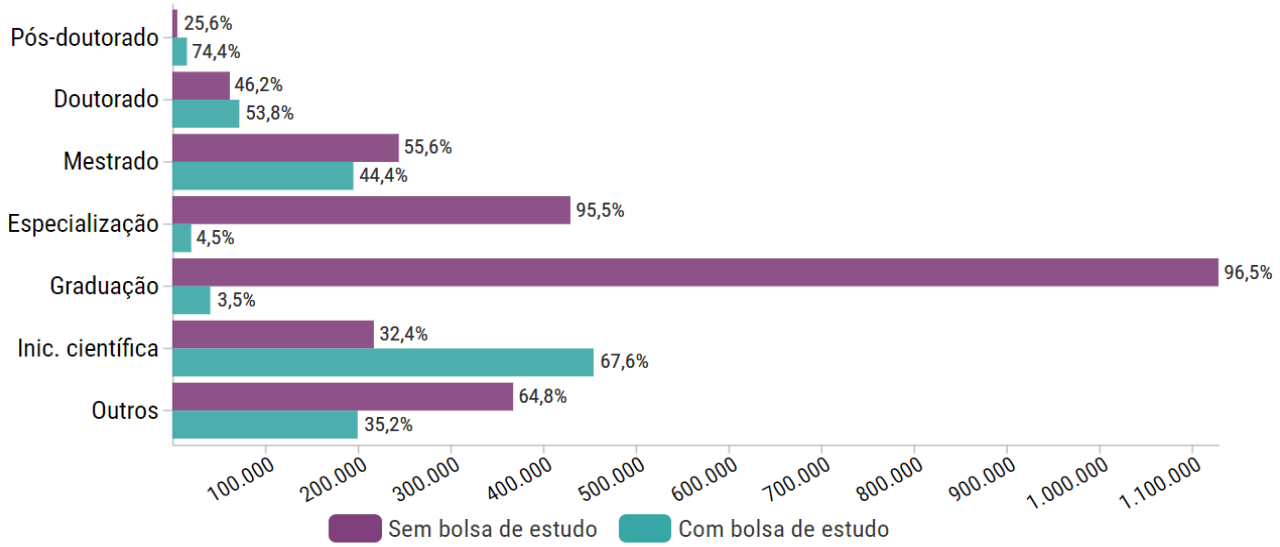
Um aspecto relevante na figura 2 refere-se a formação da pós-graduação: mestrado, doutorado e pós-doutorado. Nestas categorias, a soma das porcentagens para o gráfico de orientações concluídas corresponde a 18%, enquanto para o gráfico das orientações em andamento essa soma corresponde a 46%. Uma das hipóteses para essa diferença de percentual diz respeito ao fato de que, como os doutores são responsáveis pela formação dos alunos nos principais programas de pós-graduação stricto sensu no Brasil, eles tendem a orientar mais alunos de graduação e menos alunos de pós-graduação no início de sua carreira e, com o passar dos anos, após a conclusão do doutorado, o número de orientações da graduação diminui e as orientações da pós-graduação aumentam. No entanto, como são consideradas todas as orientações em todo o seu histórico, as orientações concluídas possuem o nível mais baixo, como graduação de formação, sendo mais representativas.

Nas orientações concluídas para a formação do pós-doutorado do RCDP, apenas 7.167 doutoras realizaram a supervisão de 21.235 pós-doutores e as duas primeiras conclusões de Pós-doutorado ocorreram em 1980, sendo supervisionadas por uma única doutora. Quanto à pós-graduação Doutorado, 22.342 doutoras orientaram 134.359 doutores e as duas primeiras conclusões ocorreram em 1969, sendo orientadas por uma única doutora. Já para a formação Mestrado, 45.593 doutoras orientaram 439.864 mestres. A primeira orientação de mestrado foi realizada em 1966. Para a Especialização, 45.609 doutoras do RCDP orientaram 450.036 especialistas e as duas primeiras orientações de especialização foram realizadas em 1966. Para a graduação, 73.609 doutoras orientaram 1.169.858 graduados e a primeira orientação realizada foi no ano de 1968. O nível Iniciação científica 60.981 doutoras orientaram 672.106 alunos. As três primeiras orientações foram realizadas no ano de 1962. Por fim, para a categoria Outros[1], 45.283 doutoras do RCDP realizaram a orientação de 567.771 orientandos e as duas primeiras orientações realizadas em 1962.

Ao preencher as informações quanto à orientação, é informado pela doutora se o orientado possuiu algum auxílio de bolsa de estudo e definir qual a agência financiadora. Assim, sumarizando esta informação, foi possível agrupar se o orientado possuiu bolsas de estudo ou não, para cada categoria referente às orientações concluídas (figura 3).

[1] As orientações concluídas podem ser cadastradas para sete diferentes categorias, em que Outras corresponde a 16% no universo da análise.

Figura 3 – Distribuição de bolsas de estudo quanto às orientações concluídas



Fonte: Elaboração dos autores.

É relevante observar na figura 3, como o percentual de bolsas foi significativo para as formações de iniciação científica, doutorado e pós-doutorado. Dentre as sete categorias, Graduação possui o menor percentual de bolsas de estudo, 3,5% e Pós-doutorado o maior, 74,36%.

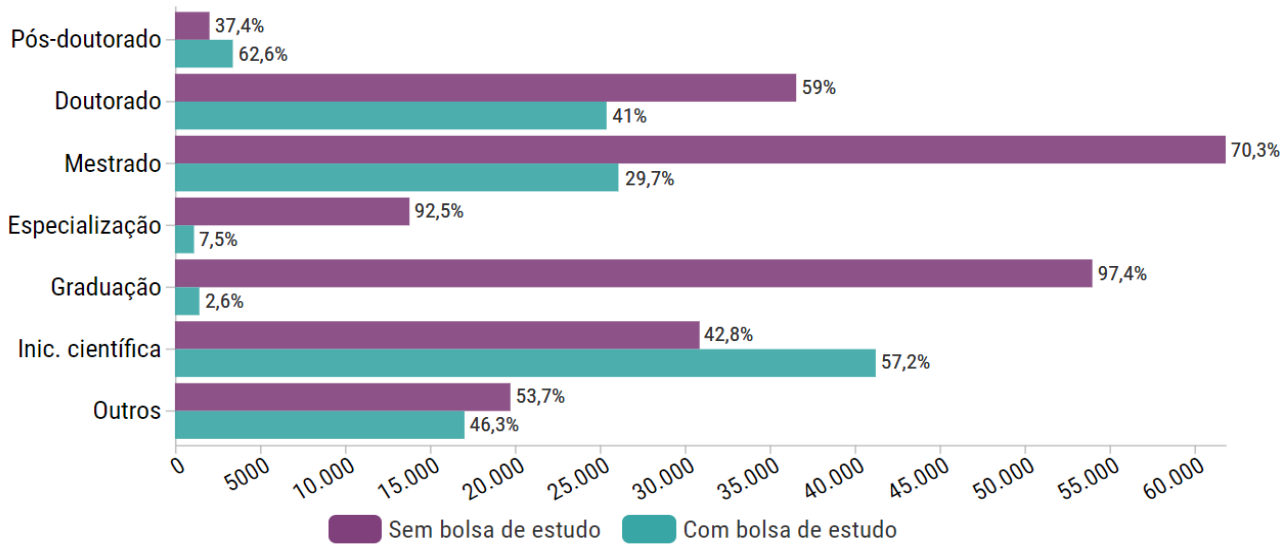
Já para as orientações em andamento para o nível de formação do pós Doutorado do RCDP, apenas 3.788 doutoras realizam a supervisão de 5.435 pós-doutorandos. Quanto ao Doutorado, 22.390 doutoras orientam 61.961 doutorandos. Já para a formação de Mestrado, 33.311 doutoras orientam 87.935 mestrandos. Para a Especialização 6.077 doutoras orientam 14.922 orientandos. Para o nível de formação Graduação, 19.699 doutoras do RCDP orientam 55.424 graduandos. O nível Iniciação científica 30.085 doutoras orientam 72.114 alunos. Por fim, para a categoria Outros, 11.193 doutoras do RCDP realizam a orientação de 36.779 orientandos.

Ao preencher as informações quanto às orientações em andamento, é informado pela doutora se o orientando possui algum auxílio de bolsa de estudo e qual a agência financiadora. Assim, sumarizando esta informação, foi possível agrupar se o orientando possui bolsas de estudo ou não, para cada categoria referente às orientações em andamento (figura 4).

É relevante observar, na figura 4, como o percentual de bolsas está sendo significativo para as formações de iniciação científica e pós-doutorado. Para mestrado e doutorado é notável o percentual de queda no total de bolsas se comparado ao percentual respectivo da figura 3. Dentre as sete categorias, a Graduação possui o menor percentual de bolsas de estudo, 2,6% e Pós-doutorado o maior, 62,6%.

Outro aspecto relevante quanto às orientações concluídas, refere-se ao atributo grande área preenchido ao inserir a orientação. Para as pós-graduações doutorado e mestrado, este item está inserido em 58,1% e 55,9%, respectivamente, já para as outras categorias, o preenchimento deste atributo é inferior a 47%. Já para as orientações em andamento, todas as categorias de formação possuem preenchimento abaixo de 37%. Reunindo as informações quanto às orientações preenchidas, foram distribuídas percentualmente, para cada categoria, as informações de orientações concluídas e em andamento de acordo com as grandes áreas (tabela 1).

Figura 4 – Distribuição de bolsas de estudo quanto às orientações em andamento



Fonte: Elaboração dos autores.

Tabela 1 – Percentual de orientações preenchidas por grandes áreas

		Pós-d.	Dout.	Mestr.	Espec.	Grad.	Inic. C.	Outros
Ciências Agrárias	Concl.	9,24%	9,47%	7,66%	2,55%	6,05%	10,18%	13,27%
	Andam.	7,02%	6,37%	5,00%	4,11%	4,49%	8,91%	10,08%
Ciências Biológicas	Concl.	22,36%	16,12%	11,00%	3,86%	6,38%	15,01%	10,96%
	Andam.	18,85%	10,18%	6,68%	3,99%	5,65%	12,17%	9,13%
Ciências da Saúde	Concl.	12,61%	19,12%	18,32%	30,11%	20,93%	19,94%	20,80%
	Andam.	14,96%	16,43%	17,09%	31,96%	20,96%	20,67%	18,41%
Ciênc. Exat. e da Terra	Concl.	10,15%	7,94%	7,17%	2,68%	4,91%	10,17%	6,62%
	Andam.	7,38%	6,77%	5,62%	2,58%	4,55%	8,83%	7,63%
Ciências Humanas	Concl.	20,70%	23,42%	26,33%	34,54%	24,98%	19,34%	22,40%
	Andam.	23,16%	29,63%	32,93%	29,74%	29,58%	20,88%	27,70%
Ciênc. Soc. Aplicadas	Concl.	6,11%	7,32%	11,90%	15,08%	24,28%	9,46%	11,09%
	Andam.	9,68%	10,56%	13,45%	14,80%	19,72%	11,34%	10,06%
Engenharias	Concl.	6,61%	5,68%	6,66%	1,98%	4,50%	6,87%	4,09%
	Andam.	5,53%	5,64%	5,42%	1,69%	4,96%	6,28%	3,52%
Ling. Letras e Artes	Concl.	11,97%	10,62%	10,34%	8,64%	7,60%	8,63%	10,26%
	Andam.	12,91%	13,74%	12,56%	10,34%	9,67%	10,32%	12,12%
Outros	Concl.	0,25%	0,31%	0,62%	0,55%	0,37%	0,38%	0,52%
	Andam.	0,51%	0,68%	1,26%	0,79%	0,41%	0,62%	1,35%
Total	Concl.	9.499	78.058	245.987	194.785	482.497	316.508	198.918
	Andam.	1.952	22.575	30.961	5.162	17.056	22.556	10.848

Fonte: Elaboração dos autores.

Assim, como pode ser observado na tabela 1, para as orientações concluídas, a categoria Pós-doutorado possui 9.499 orientações preenchidas e as duas grandes áreas mais representativas correspondem a Ciências Biológicas (2.124) e Ciências Humanas (1.966). Para as categorias Doutorado, Mestrado, Especialização e Outros, as duas grandes áreas mais representativas correspondem a Ciências Humanas e Ciências da Saúde. Já na Graduação, as duas grandes áreas mais representativas são Ciências Humanas e Ciências Sociais Aplicadas. Por fim, as duas grandes áreas mais representativas para a Iniciação Científica são Ciências da Saúde e Ciências Humanas. Já para as orientações em andamento, a formação Pós-doutorado, com 1.952 orientações em andamento preenchidas, às duas grandes áreas mais representativas correspondem a Ciências Humanas e Ciências Biológicas. Para as outras categorias, as duas grandes áreas mais representativas correspondem a Ciências Humanas e Ciências da Saúde, alterando a posição no nível Especialização.

A construção do conhecimento não ocorre de maneira isolada, é um processo interativo entre orientador/orientando, que resulta em benefícios para ambos. Conforme enunciado por Ferreira, Furtado e Silveira (2009), os benefícios para o orientando são inúmeros, dentre eles está o crescimento pessoal, profissional e acadêmico, estímulo, direção, amadurecimento, aperfeiçoamento do senso crítico, autonomia e autoconfiança. Já para o orientador os benefícios também são inumeráveis, dentre eles estão a ampliação da satisfação pessoal, incentivo e oportunidade para continuar atualizado, possibilidade para atrair novos colaboradores para projetos atuais e futuros, bem como a oportunidade de adquirir e, posteriormente, disseminar todo seu conhecimento para as gerações futuras na linha de pesquisa, como por exemplo, através das produções científicas.

Neste contexto, a principal função do orientador compreende direcionar os orientandos ao longo de sua trajetória acadêmica, mantendo "relações singulares, intersubjetivas, complexas e ricas em detalhes com os orientandos, e, desta convivência, resultam dissertações e teses que contribuem para a sistematização e consolidação do conhecimento científico em determinada área" (LEITE FILHO; MARTINS, 2006, p. 100). Para as formações mestrado e doutorado, é possível incluir orientador principal e coorientador. Assim, foi realizada a sumarização para estas duas formações, mestrado e doutorado, quanto às orientações concluídas e em andamento na figura 5.

Figura 5 – Orientador principal e coorientador: mestrado e doutorado

Fonte: Elaboração dos autores.

Para as orientações concluídas, a porcentagem referente ao orientador principal corresponde a 78,56% para doutorado e 82,69% para o mestrado. Já a porcentagem para as orientações em andamento, corresponde a 81,07% para o dourado e 85,75% para o mestrado. É importante ressaltar que no decorrer do mestrado e doutorado é possível alterar o orientador principal e incluir coorientador.

Ao preencher as orientações na Plataforma Lattes, a doutora informa o ano de término do mestrado ou doutorado para as orientações concluídas e o ano de início do mestrado ou doutorado para as orientações em andamento. Deste modo, como foram recuperados todos os anos preenchidos nas orientações concluídas, foi realizado uma sumarização, totalizando as orientações concluídas por ano e o percentual de crescimento respectivo (tabela 2).

Tabela 2 − Crescimento das orientações concluídas

		Pós-dout.	Doutorado	Mestrado	Especial.	Graduação	Inic. Cient.	Outros
2009	Quant. Cresc.	974 21,75%	6.341 10,07%	20.950 9,70%	28.535 6,28%	67.743 -1,04%	34.940 6,09%	31.364 7,81%
2010	Quant. Cresc.	1.227 25,98%	6.597 4.04%	22.153 5,74%	30.522 6,96%	71.437 5,45%	40.017 14,53%	37.159 18,48%
2011	Quant. Cresc.	1.343 9,45%	7.288 10,47%	24.389 10,09%	33.757 10,60%	72.892 2,04%	43.707 9,22%	40.600 9,26%
2012	Quant. Cresc.	1.601 19,21%	8.448 15,92%	26.843 10,06%	31.759 -5,92%	66.332 -9,00%	47.144 7,86%	41.534 2,30%
2013	Quant. Cresc.	1.869 16,74%	9.360 10,80%	29.020 8,11%	26.177 -17,58%	75.179 13,34%	48.962 3,86%	42.949 3,41%
2014	Quant. Cresc.	2.134 14,18%	10.608 13,33%	30.648 5,61%	28.042 7,12%	78.192 4,01%	48.433 -1,08%	45.434 5,79%
2015	Quant. Cresc.	2.337 9,51%	10.827 2,06%	32.863 7,23%	28.280 0,85%	72.561 -7,20%	47.919 -1,06%	43.256 -4,79%
2016	Quant. Cresc.	2.291 -1,97%	10.044 -7,23%	34.430 4,77%	24.201 -14,42%	77.931 7,40%	47.618 -0,63%	43.688 1,00%
2017	Quant. Cresc.	2.059 -10,13%	9.772 -2,71%	31.609 -8,19%	18.478 -23,65%	80.772 3,65%	45.300 -4,87%	42.107 -3,62%
2018	Quant. Cresc.	1.738 -15,59	9.416 -3,64%	25.741 -18,56%	16.600 -10,16%	70.194 -13,10%	30.954 -31,67%	33.754 -19,84%

Fonte: Elaboração dos autores.

Assim, a tabela 2 apresenta o decorrer de dez anos para todas as categorias de formação referente à taxa de crescimento anual, iniciando em 2008 e finalizando em 2018. Não foi incluído o ano de 2019, pois os dados para o estudo foram coletados neste ano e como a quantidade para cada categoria de formação estava abaixo de 50% com relação a 2018 os mesmos poderiam não ter sido incluídos por todas as doutoras e assim não são relevantes para o estudo. Nos anos 2010 e 2011, todas as categorias de formação tiveram crescimento positivo em relação ao ano anterior, enquanto que nos outros anos, pelo menos um ano de algum nível teve queda. O ano de 2018 é o único ano em que a quantidade de orientações teve queda de crescimento para todas as formações. As categorias Pós-doutorado e Doutorado tiveram crescimento quanto às orientações de 2009 até 2015, e Mestrado um ano a mais, até 2016. Ou seja, aquelas relacionadas diretamente com o desenvolvimento científico, como uma produção científica. Já as atividades focadas somente no ensino, não rendem capital "puro", pois os docentes realizam atividades invisíveis e que não geram produção científica, e, consequentemente, não alcançam reconhecimento pelo trabalho realizado.

Independentemente do valor simbólico, para que ocorra a produção científica, faz-se necessário que haja uma parceria e uma relação unilateral entre orientador e orientando. O processo de orientação envolve diversas dimensões além da produtiva, como sociais e afetivas. É imprescindível que ambas as partes se mantenham em constante diálogo e em fácil comunicação, para que possam aprender, evoluir e construir em conjunto. Porém, ressaltando o processo produtivo, é fundamental que, o orientando seja receptivo ao receber orientações e aconselhamentos para pesquisas e leituras, e simultaneamente, o orientador seja capacitado para orientar e esteja em constante aprimoramento de suas habilidades como docente (LOPES et al., 2020), podendo os dois em conjunto desenvolver o projeto e tenham bons frutos do mesmo. Ao refletirmos sobre o processo de produção científica, a orientação pode ser considerada um importante aspecto para o meio acadêmico.

As produções bibliográficas e técnicas são partes integrantes deste processo de produção de conhecimento científico, o qual possui como elementos os artigos em periódicos, livros, capítulos de livros, anais de congressos, resumos, teses, dissertações e monografias, entre outros meios de divulgação e comunicação da ciência (DOMINGUES, 2014). Como o conjunto de doutores é responsável pela maioria das produções científicas cadastradas na Plataforma Lattes (DIAS, 2016), realizou-se um levantamento das produções bibliográficas e técnicas das mulheres. Assim, dentre todos os tipos de produções disponibilizadas, foram selecionadas oito que possuem maior relevância e participação das doutoras e destes realizou-se uma sumarização do conjunto RCDP (tabela 3).

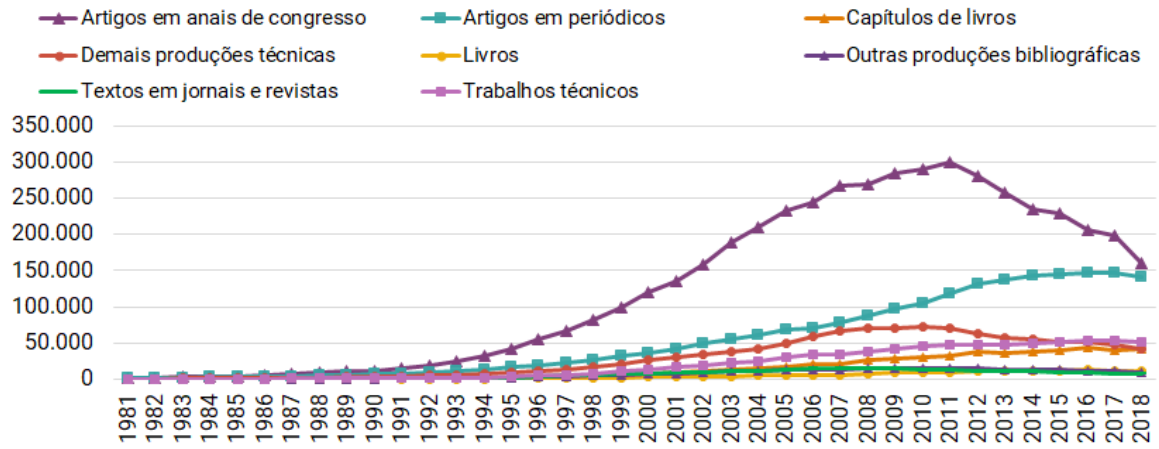
Tabela 3 – Quantitativo para as produções bibliográficas e técnicas das doutoras

Tipo de produção	Total
Artigos em Periódicos	2.155.370
Artigos em Anais de Congresso	4.811.242
Capítulos de Livros	571.110
Demais Produções Técnicas	1.262.293
Livros	184.343
Outras Produções Bibliográficas	282.503
Textos em Jornais e Revistas	269.718
Trabalhos Técnicos	807.089

Fonte: Elaboração dos autores.

As publicações podem ocorrer de diversas formas, porém, para o conjunto RCDP, a publicação e a divulgação de conteúdo se concentram nos artigos em periódicos e nos artigos em anais de congresso. Livros e capítulos de livros possuem uma quantidade bem inferior se comparado aos artigos. A partir desses dados e, levando em consideração a produtividade das doutoras por ano, foi realizada uma sumarização (figura 6) destes tipos de produções.

Figura 6 – Sumarização das produções bibliográficas e técnicas por ano



Fonte: Elaboração dos autores.

Para esta análise, foram contabilizados todos os trabalhos registrados na Plataforma Lattes por currículo do RCDP, ou seja, o mesmo trabalho registrado em currículos distintos foi contabilizado em sua totalidade. A primeira publicação registrada no RCDP para artigos em periódicos ocorreu em 1940 e anais de congressos em 1948. Neste cenário de publicações por ano, é possível perceber que, nos primeiros quarenta e um anos (1940 a 1981), a produção bibliográfica permaneceu tímida e constante para todos os tipos de produções. Possivelmente, uma hipótese para explicar esse fato pode ser relacionada ao lançamento e padronização do currículo da Plataforma Lattes que ocorreu em agosto de 1999, por essa razão, os dados referentes às publicações anteriores a este período podem não ter sido divulgados pelas doutoras. Outra hipótese que pode explicar esse fato está relacionada ao número de doutoras, que eram poucas na época.

A partir do início da década de 1980, ocorreu um aumento em todos os tipos de produções bibliográficas e tecnológicas. As produções que se evidenciaram sobre as demais correspondem aos artigos em periódicos e em anais de congressos. Com relação aos artigos em periódicos, apresentou-se um crescimento significativo até 2013, permanecendo constante após esse ano e em queda em 2017.

Já as produções dos artigos em anais e congressos, tiveram um aumento considerável com ápice no ano de 2011 e queda significativa após esse ano. Esse declive acentuado referente foi tão expressivo que desde o ápice até o ano de 2018, apresentou uma queda de 46,68%, chegando ao final do último ano com valor total de artigos próximo ao total dos artigos em periódicos. Essa diminuição considerável no número de artigos em anais de congresso apresenta o mesmo comportamento no estudo de Dias (2016), para o total de publicações por ano referente ao conjunto de todos os doutores com currículos cadastrados na Plataforma Lattes.

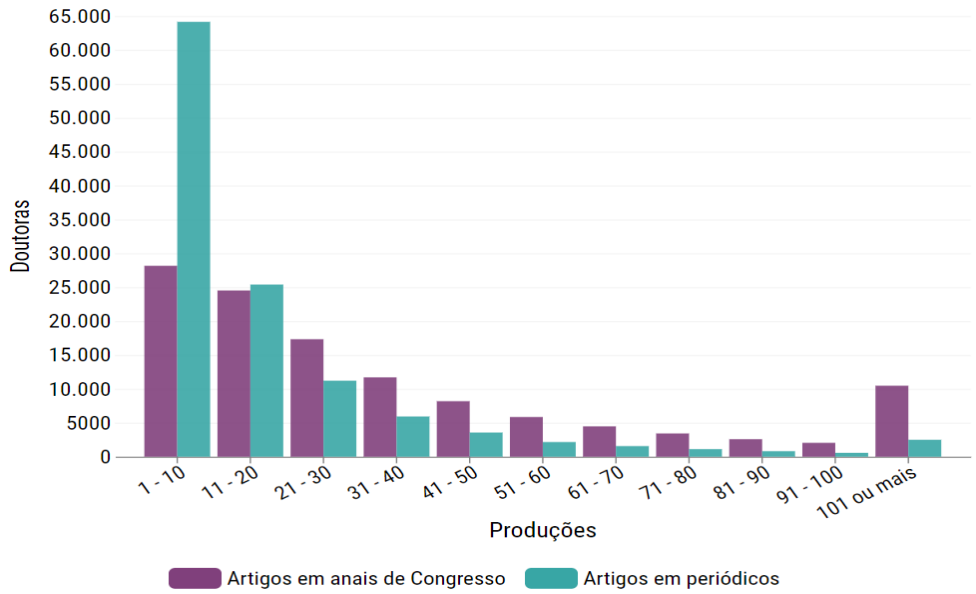
São diferentes hipóteses que podem estar relacionadas ao declive acentuado dos artigos em anais de congresso a partir de 2011. Uma delas refere-se à classificação da produção científica utilizada pela CAPES, se considerarmos que o sistema de avaliação da produção influencia as ações dos indivíduos. Fato esse impulsionado pela não consideração dos artigos em anais de congresso nas avaliações dos programas de pós-graduação. Assim, os artigos em periódicos tornam-se mais interessantes para a publicação, pois influenciam nos conceitos dos programas do qual os doutores participam, e os mesmos direcionam seus esforços para esse tipo de publicação.

Atualmente, a "[...] classificação dos periódicos é organizada através do módulo Qualis da Plataforma Sucupira, que coleta dados para avaliação dos programas de pós-graduação recomendados pela Capes" (FREIRE; FREIRE, 2017, p. 5). O Qualis-Periódicos é "um sistema usado para classificar a produção científica dos programas de pós-graduação no que se refere aos artigos publicados em periódicos científicos" (SUCUPIRA, 2020), aferindo a qualidade da produção a partir da análise da qualidade dos veículos de divulgação. A avaliação é periódica e possui as classificações de 2010-2012 e 2013-2016. Possui indicativos de qualidade A1, mais elevado; A2; B1; B2; B3; B4; B5; C - peso zero.

Continuando com a análise das produções bibliográficas das doutoras e objetivando demonstrar a proporção de doutoras de acordo com seus números de produções, foi recuperado o total de artigos de cada doutora e distribuído em períodos. Assim, selecionaram-se dez períodos contendo um intervalo de 10 anos em cada e, para o último período, agruparam-se as doutoras que possuem mais de 101 artigos, apresentando então no eixo vertical a quantidade de doutoras e no eixo horizontal o número de produções, para artigos em periódicos e em anais de congresso (figura 7).

O número de doutoras que não possuem produções corresponde a 6.509 (5,19%) para artigos em periódicos e 6.720 (5,35%) para artigos em anais de congresso. Já referente ao último período, que corresponde às doutoras que possuem mais de 101 artigos, é importante ressaltar que 13 doutoras possuem cada uma mais de 500 artigos produzidos e a maior quantidade corresponde a 737 para os artigos em periódicos. Já para os artigos em anais de congresso, 17 doutoras possuem mais de 750 artigos produzidos, sendo a maior produção correspondendo a 1.827 artigos. Ao observarmos a figura 7, o primeiro período possui muita representação para os artigos em periódicos ao contemplar um total de 51,11%, ou seja, mais da metade das doutoras possui total de produções entre 1 até 10 artigos, que corresponde a um número baixo de produção para uma grande quantidade de doutoras, se comparado aos outros períodos. O segundo período possui pouco menos da metade do período anterior, que corresponde a 20,24%, ocorrendo sucessivamente para os outros períodos.

Figura 7 – Quantitativo das produções por doutoras



Fonte: Elaboração dos autores.

O período que corresponde a 101 ou mais, compreende um total de 688.543 artigos em periódicos para 2.498 (1,99%) doutoras, ou seja, possui uma quantidade significativa de artigos para um pequeno número de doutoras. Já para os artigos em anais de congresso, os três primeiros períodos contemplam um total de 55,79%, ou seja, mais da metade das doutoras produziram de 1 até 30 artigos. O período que corresponde a 101 ou mais, compreende um total de 1.772.670 artigos em anais de congresso para 10.474 (8,34%) doutoras, ou seja, possui uma quantidade significativa de produções para um pequeno número de doutoras. Essa análise também foi realizada para livros e capítulos de livros. Um número considerável de doutoras não possui produções, correspondendo a 77.540 (61,78%) para livros e 45.805 (36,49%) para capítulos de livros. Para as produções em livros, o período mais representativo corresponde ao primeiro, de 1 a 10 livros publicados, totalizando 44.597 (35,53%) doutoras. São 12 (0,01%) doutoras agrupadas no período 101 ou mais, sendo a maior produção de 134 livros. Já para os capítulos de livros, o primeiro período também é o mais significativo, compreendendo 64.613 (51,48%) doutoras. São 89 (0,07%) doutoras agrupadas no período 101 ou mais, sendo que a maior produção corresponde a 333.

É importante ressaltar que 13 doutoras hiper produtivas possuem cada uma mais de 500 artigos em periódicos, sendo que a maior produção corresponde a 737 artigos. Já para artigos em anais de congresso, 17 doutoras hiper produtivas possuem cada uma mais de 750 artigos produzidos, sendo que a maior produção corresponde a 1.827. Duas doutoras são hiper produtivas, tanto para artigos em anais de congresso, quanto para artigos em periódicos e possuem a mesma grande área de Ciências da Saúde, são elas: Vera Luiza Capelozzi e Angela Maggio da Fonseca. As outras doutoras hiper produtivas possuem as seguintes grandes áreas: Ciências da saúde 14, Ciências exatas e da terra 6, Ciências agrárias 2, Ciências biológicas 2, Engenharias 2, Ciências humanas 1, não informado 1. Um estudo destacando as hiper produtivas pode ser observado por Tuesta (2019), no qual 5% das mulheres publicaram mais de 130 artigos.

Outro dado relevante corresponde à data de atualização dos currículos do RCDP, em que 69,05% das doutoras atualizaram seus currículos em 2019. Ao relacionar a figura 7 com as datas de atualizações dos currículos, foi possível mapear e distribuir o percentual de atualização por ano de acordo com cada período (tabela 4).

Tabela 4 − Atualização dos currículos de acordo com a quantidade de produção

	2000 a 2004	2005 a 2009	2010 a 2014	2015	2016	2017	2018	2019
0	0,54%	3,56%	14,63%	5,52%	6,78%	8,96%	18,79%	41,24%
1 a 10	0,22%	1,64%	5,82%	2,53%	3,71%	6,44%	16,20%	63,44%
11 a 20	0,17%	1,25%	4,11%	1,59%	2,40%	4,33%	11,29%	74,86%
21 a 30	0,12%	1,11%	4,00%	1,42%	1,85%	3,36%	8,30%	79,84%
31 a 40	0,20%	0,94%	3,54%	1,21%	1,90%	2,76%	7,65%	81,79%
41 a 50	0,20%	1,12%	2,78%	1,43%	1,74%	2,44%	6,03%	84,27%
51 a 60	0,05%	1,29%	3,60%	1,15%	1,85%	2,40%	6,14%	83,51%
61 a 70	0,06%	1,15%	3,95%	1,28%	1,47%	2,36%	6,63%	83,10%
71 a 80	0,18%	0,80%	2,68%	1,70%	0,89%	2,68%	6,25%	84,82%
81 a 90	0	0,86%	2,44%	0,98%	0,98%	1,83%	5,13%	87,78%
91 a 100	0,52%	0,70%	2,61%	1,57%	1,92%	1,92%	4,88%	85,89%
101 ou mais	0,12%	0,76%	2,36%	0,92%	1,40%	1,76%	5,04%	87,63%

Fonte: Elaboração dos autores.

Observando a tabela 4, para as doutoras que não possuem artigos verifica-se o percentual de atualização baixo, se comparado aos outros períodos. Uma hipótese para este fato pode ser relacionada a não atuação das doutoras no ensino e pesquisa. Como um dos requisitos ao participar de processos seletivos ou mesmo submeter projetos nos editais é ter informação atualizada no currículo, caso estas doutoras não estiverem atuando no ensino e pesquisa, não têm a necessidade de realizar periodicamente esta atualização. Já os três últimos períodos, nos quais as doutoras produziram mais de 81 artigos, possuem a data de atualização com os maiores percentuais para o ano de 2019. Assim, sendo a produção científica uma importante atividade acadêmica, na qual as pesquisadoras são ativas e atualizam seus currículos constantemente, principalmente as doutoras que publicam uma quantidade elevada de artigos. Conforme Dias e Moita (2018) argumentam, currículos que possuem a data de atualização mais recente provavelmente já possuem trabalhos recém-publicados registrados, o que proporciona uma visão atual da produção científica brasileira.

CONSIDERAÇÕES FINAIS

As mulheres nas ciências têm sido foco de inúmeros estudos e este abordou a perspectiva da orientação e da produção bibliográfica e técnica das doutoras pesquisadoras. Utilizando como fonte de dados as bases da Plataforma Lattes, foi possível selecionar, pelos critérios de formação/titulação doutorado concluído e de gênero feminino, as doutoras que já participaram ou estão atualmente integradas a um grupo de pesquisa. Assim, a preparação e análise dos dados foi realizada utilizando o RCDP, que contempla 125.515 currículos das doutoras pesquisadoras.

As orientações correspondem a um tópico relevante, pois ocorrem em diversos níveis de escolaridade, em que o conhecimento é construído através do processo interativo entre o orientador e o orientando. Assim, foi apresentado o perfil das orientações das doutoras pesquisadoras, como as orientações concluídas, as orientações em andamento, orientação principal e coorientação para o mestrado e doutorado, crescimento das orientações para um período de 10 anos.

O processo de produção científica é o fruto de uma investigação científica e se enquadra em uma das diversas atividades de pesquisa e ensino do docente-pesquisador. Como os doutores são responsáveis pela maior parte das produções científicas na Plataforma Lattes, este é um tópico expressivo para o estudo. Assim, foi apresentada a representatividade da produção bibliográfica e técnica das doutoras, como o total de produções bibliográficas e técnicas, a sumarização das produções por ano, distribuição das produções por período, atualização dos currículos por período de acordo com a quantidade de produção.

Os estudos que tem como temática os aspectos gerais da mulher e particularmente na ciência, são relevantes e podem contribuir para gerar indicadores científicos nacionais e para a gestão das informações na área científica e tecnológica, e este apresentou dois aspectos da mulher na ciência sobre o conjunto de dados. Como trabalhos futuros, espera-se realizar uma análise da produção bibliográfica das pesquisadoras que são bolsistas de produtividade em pesquisa do CNPq e a colaboração científica das doutoras pesquisadoras.

REFERÊNCIAS

ARAÚJO, C. A. Bibliometria: evolução histórica e questões atuais. *Em Questão*, Porto Alegre, v. 12, n. 1, p. 11–32, jan./jun. 2006. Disponível em: https://seer.ufrgs.br/EmQuestao/article/view/16. Acesso em: mar. 2021.

BOURDIEU, P. *Usos sociais da ciência*. São Paulo: Editora Unesp, 2004.

DIAS, T. M. R. *Um Estudo da Produção Científica Brasileira a partir de Dados da Plataforma Lattes*. 2016. 181 p. Tese (Doutorado em Modelagem Matemática e Computacional) — Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, set. 2016.

DIAS, T. M. R.; MOITA, G. F. Um retrato da produção científica brasileira baseado em dados da plataforma lattes. *Brazilian Journal of Information Science*: research trends, Marília, SP, v. 12, n. 4, p. 62–74, 2018. DOI: https://doi.org/10.36311/1981-1640.2018.v12n4.08.p62. Disponível em: https://revistas.marilia.unesp.br/index.php/bjis/article/view/7831. Acesso em: mar. 2021.

DIRETÓRIO DE GRUPOS DE PESQUISA NO BRASIL. *Censos do Diretório de Grupos de Pesquisa*. Brasília: CNPq, 2019. Disponível em: http://lattes.cnpq.br/web/dgp/censos2. Acesso em: 1 out. 2019.

DOMINGUES, I. O sistema de comunicação da ciência e o taylorismo acadêmico: questionamentos e alternativas. *Estudos avançados*, São Paulo, v. 28, n. 82, p. 225–250, dez. 2014. DOI: https://doi.org/10.1590/S0103-40142014000300014. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-40142014000300014&lng=en&nrm=iso. Acesso em: mar. 2021.

FERREIRA, L. M. FURTADO, F.; SILVEIRA, T. S. Advisor-advisee relationship: the multiplier knowledge. *Acta Cirúrgica Brasileira*, São Paulo, v. 24, n. 3, p. 170–172, jun. 2009. DOI: https://doi.org/10.1590/S0102-86502009000300001. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-86502009000300001&lng=en&nrm=iso. Acesso em: mar. 2021.

FREIRE, G. H. A.; FREIRE, I. M. Sobre o qualis de periódicos da capes. *Inf. & Soc.:Est.*, João Pessoa, v.27, n.3, p. 5-6, set./dez. 2017. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/37560. Acesso em: mar. 2021.

LANE, J. Let's make science metrics more scientific. *Nature*, v. 464, n. 7288, p. 488-489, 25 mar. 2010. Disponível em: https://doi.org/10.1038/464488a. Acesso em: 27 mar. 2019.

LEITE FILHO, G. A.; MARTINS, G. A. Relação orientador-orientando e suas influências na elaboração de teses e dissertações. *Revista de Administração de Empresas,* São Paulo, v. 46, n. esp., p. 99–109, dez. 2006. DOI: https://doi.org/10.1590/S0034-75902006000500008. Disponível: https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-75902006000500008&lng=pt&tlng=pt. Acesso em: mar. 2021.

LETA, J. Mulheres na ciência brasileira: desempenho inferior? *Revista Feminismos*, v. 2, n. 3, p. 139–152, set./dez. 2014. Disponível em: https://periodicos.ufba.br/index.php/feminismos/article/view/30039/17771. Acesso em: mar. 2021.

LOPES, E. F. B. *et al.* A relação entre orientador e orientando no processo de produção científica = The relationship between guiding and guiding in the scientific production process. *Brazilian Journal of Development*, v. 6, n. 1, p. 3854–3868, jan. 2020. DOI:10.34117/bjdv6n1-273. Disponível em: https://www.brazilianjournals.com/index.php/BRJD/article/view/6352/5630. Acesso em: mar. 2021.

SUCUPIRA. *Qualis*. Brasil, 2020. Disponível em: https://sucupira.capes.gov.br/sucupira/public/index.xhtml#. Acesso em: 7 abr. 2020.

TUESTA, E. F. *et al.* Análise da participação das mulheres na ciência: um estudo de caso da área de Ciências exatas e da Terra no Brasil. *Em Questão*, Porto Alegre, v. 25, n. 1, p. 37–62, jan./abr. 2019. DOI: https://doi.org/10.19132/1808-5245251.37-62. Disponível em: https://seer.ufrgs.br/EmQuestao/article/view/80193. Acesso em: mar. 2021.

AGRADECIMENTOS

Os autores agradecem ao CEFET-MG e CAPES pelo auxílio na pesquisa.

A publicidade de dados abertos pelo Tribunal Superior Eleitoral (TSE): o caso do Repositório de Dados Eleitorais

Márcio Bezerra da Silva
Doutor em Ciência da Informação pela Universidade Federal da Bahia (UFBA) - Salvador, BA - Brasil. Professor da Faculdade de Ciência da Informação (FCI) da Universidade de Brasília (UnB) - Brasília, DF - Brasil.
http://lattes.cnpq.br/9275164094039775
E-mail: marciobdsilva@unb.br

Rafael Fernandes de Barros Costa Azevedo
Mestre em Ciência da Informação pela Universidade de Brasília (UnB) - Brasília, DF - Brasil.
Coordenador de Logística da Secretaria de Tecnologia da Informação do Tribunal Superior Eleitoral (TSE) - Brasília, DF - Brasil.
http://lattes.cnpq.br/7269157765737686
E-mail: rafaelfbca@gmail.com

Denise de Oliveira Araújo
Graduada em Biblioteconomia pela Universidade de Brasília (UnB) - Brasília, DF - Brasil.
Auxiliar de biblioteca, atua no Serviço de Gestão do Conhecimento (SGCo) do Tribunal de Contas da União (TCU), Brasília, DF - Brasil.
http://lattes.cnpq.br/5721499163118225
E-mail: deoliveiraraujo@gmail.com

Fernanda Percia França
Graduada em Biblioteconomia pela Universidade de Brasília (UnB) - Brasília, DF - Brasil.
Biblioteca do Senado Federal - Brasília, DF - Brasil.
http://lattes.cnpq.br/5003950145456243
E-mail: fernanda10nov@hotmail.com

Marilete da Silva Pereira
Graduanda em Biblioteconomia pela Universidade de Brasília (UnB) - Brasília, DF - Brasil.
Aluna do Programa Institucional de Bolsas de Iniciação Cientifica (PIBIC) - Brasil.
http://lattes.cnpq.br/6950192919515947
E-mail: mariletesilvaunb@gmail.com

Submetido em: 30/04/2020. Aprovado em: 21/05/2021. Publicado em: 28/07/2021 .

RESUMO

Este artigo apresenta o Repositório de Dados Eleitorais, do Tribunal Superior Eleitoral, no contexto dos dados abertos governamentais. Ele se fundamenta na literatura sobre dados abertos, governo aberto e dados abertos no Governo Federal e na Justiça Eleitoral brasileira. Adotando o Repositório de Dados Eleitorais enquanto objeto de estudo, investigado como uma pesquisa descritiva, por meio de técnicas bibliográficas e pesquisa-ação para a coleta de dados em sentido qualitativo, este trabalho apresenta um repositório resultante de um processo, natural de informatização do processo eleitoral, criado por um setor específico de tecnologia da informação e gerido, atualmente, pelo Núcleo de Estatística, além de atualizado periodicamente, respeitando as retotalizações. Essa compilação de dados oferta informações sobre candidatos, abstenção, eleitorado, partidos, pesquisas eleitorais, prestação de contas, resultados etc., enquanto dados eleitorais abertos, coletados por um Extract, Transform, Load e disponibilizados como dados brutos, de maneira compactada, contendo, em seu interior, um arquivo com os próprios dados eleitorais em Comma-Separated Values e outro, em Portable Document Format, com a descrição dos dados disponibilizados. Concluiu-se que a transparência é o fio condutor do Tribunal Superior Eleitoral, o que fomentou o desenvolvimento do Repositório de Dados Eleitorais, com fins de disponibilizar, baseado em tecnologias específicas, dados eleitorais abertos, passíveis de uso em qualquer programa do tipo planilha eletrônica e oriundos do processo eleitoral, desde o cadastramento e a gestão de eleitores até a divulgação dos votos, e, assim, contribuir para a garantia da imparcialidade e legitimidade dos pleitos eleitorais.

Palavras-chave: Dados abertos. Dados abertos governamentais. Dados eleitorais abertos. Governo aberto.

Advertising of open data by the Electoral High Court(EHC): the case of the Electoral Data Repository

ABSTRACT

This Article aims to present the Electoral Data Repository, of the Superior Electoral Court, in the context of open government data. It is based on literary texts on open data, open government and open data on the Federal Government and the Brazilian Electoral Justice. Adopting the Electoral Data Repository as the object of study, investigated as a descriptive research, using bibliographic techniques and action research for data collection, in a qualitative sense, this work presents a repository resulting from a natural process of computerization of the electoral process, created by a specific sector of information technology and currently managed by the Statistics Center and updated periodically, respecting retotalizations. This data compilation offers information on candidates, abstention, electorate, parties, electoral polls, accountability, results, etc., while open electoral data collected by an Extract, Transform, Load and made available, as raw data, of compressed way, containing, inside, a file with the electoral data themselves, in Comma-Separated Values, and another in Portable Document Format, with the description of the available data. Concluded that transparency is the guiding thread of the Superior Electoral Court, which fostered the development of the Electoral Data Repository, with the purpose of making available, based on specific technologies, open electoral data, which can be used in any electronic spreadsheet program and coming from the electoral process, from voter registration and management to the dissemination of votes, and thus contribute to ensuring the impartiality and legitimacy of electoral elections.

***Keywords**: Open data. Open government data. Open electoral data. Open government.*

Publicidad de datos abiertos por el Tribunal Superior Electoral(TSE): el caso del depósito de datos electorales

RESUMEN

Artículo tiene como objetivo presentar el Depósito de datos electorales, del Tribunal Superior Electoral, en el contexto de los datos abiertos del gobierno. Se basa en textos literarios sobre datos abiertos, gobierno abierto y datos abiertos en el Gobierno Federal y en la Justicia Electoral brasileña. Adoptando el repositorio de datos electorales como objeto de estudio, investigado como una investigación descriptiva, utilizando técnicas bibliográficas e investigación de acción para la recopilación de datos, en un sentido cualitativo, este trabajo presenta un repositorio resultante de un proceso natural de informatización del proceso electoral, creado por un sector específico de tecnología de la información y actualmente administrado por el Centro de Estadísticas y actualizado periódicamente, respetando las retotalizaciones. Esta recopilación de datos ofrece información sobre candidatos, abstención, electorado, partidos, encuestas electorales, rendición de cuentas, resultados, etc., mientras se abren datos electorales recopilados por un Extracto, Transformación, Carga y se ponen a disposición, como datos brutos, de manera compacta, que contiene, en su dentro, un archivo con los datos electorales en sí, en valores separados por comas, y otro en formato de documento portátil, con la descripción de los datos proporcionados. Se concluyó que la transparencia es el hilo conductor del Tribunal Superior Electoral, que promovió el desarrollo del Depósito de Datos Electorales, con el propósito de poner a disposición, con base en tecnologías específicas, datos electorales abiertos, que se pueden utilizar en cualquier programa de hoja de cálculo electrónica. y desde el proceso electoral, desde el registro y la gestión de los votantes hasta la difusión de los votos, y así contribuir a garantizar la imparcialidad y la legitimidad de las elecciones electorales.

Palabras clave*: Datos abiertos. Datos gubernamentales abiertos. Datos electorales abiertos. Gobierno abierto.*

INTRODUÇÃO

A crescente atividade de produção e compartilhamento de dados vincula-se à evolução das Tecnologias de Informação e Comunicação (TICs), as quais foram se popularizando e, com o tempo, se tornando essenciais no cotidiano dos indivíduos. A principal ferramenta desse cenário é a *web,* que se faz presente desde tarefas básicas, como enviar um *e-mail* ou assistir a um filme, até propiciar a interação do indivíduo com o governo (CONEGLIAN *et al.*, 2018).

No cenário da *web* está o grande fluxo de dados produzidos e as atividades oriundas desta produção, o que, de certa forma, possui sua parcela de contribuição ao fomento da *Data Science* (Ciência de Dados). Cardoso (2019) compreende a Ciência de Dados como um campo transdisciplinar, valendo-se da sintetização de disciplinas como: Estatística, Ciência da Computação e Ciência da Comunicação.

Ao receber contribuições de outras áreas, assim como apoia, consequentemente, outras mais, a Ciência de Dados constitui-se de

> [...] metodologias, teorias, tecnologias e aplicativos relevantes para dados, desde a captura, criação, representação, armazenamento, pesquisa, compartilhamento, privacidade, segurança, modelagem, análise, aprendizagem, apresentação e visualização, até a integração de recursos complexos, heterogêneos e interdependentes para a tomada de decisões em tempo real, colaboração, criação de valor e suporte à decisão. (CARDOSO, 2019, p. 53).

Em face do exposto, a Ciência de Dados preocupa-se com o chamado ciclo de vida dos dados, que, a partir de suas fases – coleta, armazenamento, recuperação e descarte –, e de pilares específicos – privacidade, integração, qualidade, direitos autorais, disseminação e preservação –, intenciona garantir a existência saudável dos dados (CARDOSO, 2019).

Nessa conjuntura, tendo em mente tópicos como privacidade, qualidade, direitos autorais, disseminação e preservação de dados, é possível inferir que a Ciência de Dados realiza e participa de discussões em que os dados sejam o tópico central, como os dados abertos, incluindo, por exemplo, a variação governamental.

Entre os tipos de dados estão os chamados "abertos", os quais foram, em 2009, o cerne de um movimento denominado *Open Data* (Dados Abertos), quando países, entre eles Estados Unidos da América (EUA) e Inglaterra, iniciaram um modelo de gestão com o objetivo de ampliar a visibilidade de informações governamentais, a fim de produzir efeitos que conduzissem a população a contribuir para a eficiência e a transparência dos governos, e, assim, fortalecer a participação da sociedade na gestão governamental (SANTAREM SEGUNDO, 2013).

Diante do alinhamento da intenção de abrir os dados de governo com as iniciativas relacionadas aos dados abertos, origina-se o conceito de dados abertos governamentais, ocasionando ênfase nas iniciativas de dados abertos, devido ao valor que possuem para a sociedade (BRANDT; VIDOTTI; SANTAREM SEGUNDO, 2018). No Brasil, a iniciativas de dados abertos governamentais foram impulsionadas pela Lei de Acesso à Informação (LAI), nº 12.527[1], de 2011, tornando obrigatória a divulgação de dados abertos pelos órgãos públicos brasileiros, e pelo Decreto nº 8.777[2], de 2016, que instaurou a Política de Dados Abertos (PDA) do Poder Executivo Federal, também obrigando os órgãos que compõem o Poder Executivo Federal Brasileiro a publicarem uma PDA.

Assim, torna-se fundamental que esses dados tanto sejam publicados de acordo com as melhores condutas e diretrizes das comunidades de prática da *web*, quanto estejam disponíveis em formato aberto, ou seja, em consonância com princípios que autorizem a manipulação, a reutilização e o tratamento, nesse caso, de maneira livre, desses dados, pensamento que vai ao encontro da ideia de dados abertos governamentais (*ibid*).

Na perspectiva do governo, ao citar a transparência, ainda é possível elencar a Justiça Eleitoral. Autores como Kaplan (2006) defendem que a transparência é essencial para o processo eleitoral, pois aumenta a confiabilidade do público no sistema eleitoral e limita as possibilidades de fraude. Conforme a Lei nº 8.868, de 1994, o pleito eleitoral deve ocorrer da mesma forma em todo o país, usando as mesmas regras, independentemente de o processo ser estadual ou municipal. Outro ponto em favor do conceito de transparência concerne à disponibilização de dados abertos na Internet, nesse caso, de dados eleitorais abertos, acompanhando os comportamentos induzidos pelas TICs na sociedade. Conforme o Tribunal Superior Eleitoral (TSE) ([2020?b]), devido à crescente demanda de informações solicitadas por pesquisadores, imprensa e demais pessoas interessadas em analisar dados eleitorais, como candidaturas, eleitores e resultados, foi desenvolvido o Repositório de Dados Eleitorais (RDE)[3], que disponibiliza um conjunto de informações brutas das eleições.

Ao direcionar a discussão para o TSE, enquanto espaço que compõe o Governo Federal, problematizou-se as características que definem os dados eleitorais abertos e o modo como esses dados são coletados e disponibilizados. Nesse sentido, este artigo objetiva apresentar o RDE no contexto dos dados abertos governamentais. Especificamente, almeja elencar características de sua criação e periodicidade de atualização, identificar quais sãos os tipos de dados abertos disponibilizados, citar fases do clico de vida dos dados no repositório e apontar a forma como os dados eleitorais são coletados e disponibilizados.

[1] Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm.

[2] Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm.

[3] Disponível em: http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1/repositorio-de-dados-eleitorais.

METODOLOGIA

Com o intuito de responder aos objetivos do artigo, a pesquisa baseou-se no método dedutivo, tendo o RDE como objeto de estudo, considerando o período de dezembro de 2019 a abril de 2020. Ela caracterizou-se, ainda, como pesquisa descritiva, com foco em apresentar características do objeto estudado, e qualitativa, em vista da abordagem de coleta de dados. Para tanto, adotou-se a pesquisa-ação no recolhimento de informações sobre o TSE e o RDE, pois um dos autores, rotulado no artigo como "servidor A", faz parte do Tribunal e já atuou na implementação do repositório. Assim, um questionário foi enviado ao "avaliador A" em 26 de abril de 2020 e as respostas sobre dados eleitorais, criação do RDE (período, equipe e infraestrutura), dados abertos no TSE, periodicidade de atualização e coleta e disponibilização dos dados eleitorais foram produzidas e analisadas até o dia 28 do mesmo mês. Além dos dados apresentados pelo "servidor A", foram coletadas informações no próprio RDE, ao ser estudado, permitindo identificar os tipos de dados abertos disponibilizados no repositório, como candidatos, eleitorados, partidos etc.

Optou-se, também, pela pesquisa bibliográfica em periódicos científicos identificados no Portal de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), cuja filtragem considerou os que possuem *Qualis* de valores A1, A2 ou B1, além de outras fontes, como artigos de anais de congresso, capítulos de livros, dissertações, teses, dicionários, leis, resoluções, revistas jurídicas, portais de notícias e *websites* especializados.

DADOS ABERTOS: UMA BREVE CONTEXTUALIZAÇÃO

O conceito de dados abertos (*open data*) refere-se aos dados públicos ou privados, legíveis por meio de máquinas, mas sem padrões tecnológicos, econômicos, sociais ou legais, que visam à acessibilidade a qualquer sujeito. Nesse caso, exige-se, no máximo, que se atribua à fonte original e que o compartilhamento seja realizado pela mesma licença em que os dados foram apresentados (THE OPEN DEFINITION, [2020?]).

Ou seja, deve-se evitar que mecanismos controlem ou restrinjam o acesso a dados que forem publicados *on-line*, permitindo que qualquer pessoa, física ou jurídica, os explore, acesse, modifique e compartilhe livremente, para qualquer fim, contudo, estando sujeito (o acesso), quando muito, à exigência de requisitos como compartilhar pela mesma licença e/ou citar a fonte original (GOV.BR, [2020?]; OPEN KNOWLEDGE INTERNATIONAL [2020?]).

Os dados abertos contêm três normas fundamentais: **disponibilidade e acesso**, visto que os dados devem ser disponibilizados integralmente, a um custo razoável de reprodução, *on-line* e de forma que possam ser modificados; **reutilização e redistribuição**, requerendo que os dados contenham termos que permitam o seu reúso e a sua redistribuição, inclusive, tendo a capacidade de interoperar conjuntos diferentes de dados; e **participação universal**, exigindo que os dados sejam usados, reutilizados e redistribuídos, sem discriminação de área de atuação, pessoas ou grupos (OPEN KNOWLEDGE INTERNATIONAL [2020?]).

À medida que o conceito de dados abertos se solidificava, diferentes movimentos globais aconteciam, considerando, especialmente, a transparência de dados. Isotani e Bittencourt (2015) citam: a campanha feita pela *International Aid Transparency Initiative* (IATI)[4], focada na transparência dos registros de gastos de recursos humanitários; o movimento que reuniu ativistas da Internet (2017) para conceituar dados abertos públicos ou governamentais, partindo da premissa de que, assim como as ideias científicas, os dados governamentais são de todos, ou seja, uma propriedade comum; e a *Open Government Partnership* (OGP)[5], a qual, nascida em 2011, tem o Brasil como um de seus criadores e se instituiu por intermédio da participação de 65 países.

[4] Disponível em: https://iatistandard.org/en/.

[5] Disponível em: https://www.opengovpartnership.org/.

Demonstrando que o conhecimento intercambiado beneficia a todos, pois promove autores e informações completas entre diferentes países, cita-se a área da ciência, onde, graças ao movimento de dados científicos abertos, houve a maximização da disseminação e disponibilização do conhecimento, o que possibilitou acesso a dados de pesquisa e, consequentemente, passou a modificar a comunicação científica (POSSAMAI, 2016).

Exemplo disso, é o Repositório de Dados Científicos da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), que disponibiliza dados de pesquisas gerados em universidades localizadas no estado de São Paulo. Atuando como um metabuscador de dados de pesquisa, possibilita pesquisar pelas Instituições que fazem parte da federação e explorar os conteúdos (dados) referentes a autor, assunto e data de publicação (FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO, 2019).

Do ponto de vista político, o movimento do acesso aberto, público ou governamental, teve seu fortalecimento em 2008, nos EUA, a partir da *Transparency and Open Government* (Transparência e Governo Aberto), iniciativa que apresentou as recomendações necessárias à disponibilização dos dados abertos do Governo norte-americano (ISOTANI; BITTENCOURT, 2015; POSSAMAI, 2016).

GOVERNO ABERTO

A popularização do termo Governo Aberto apenas ocorreu durante a campanha de Barack Obama (ex-presidente dos EUA), quando o massivo uso de redes sociais, na corrida presidencial, aproximando eleitores, tornou-se um marco no *marketing* político sob um viés aberto (POSSAMAI, 2016). Eleito em 2008, Obama estabeleceu a *Transparency and Open Government*, apresentando orientações para que as agências governamentais coletassem e reunissem todos os dados em um só repositório, disponibilizando-os a qualquer cidadão e empresa na *web* (ISOTANI; BITTENCOURT, 2015; POSSAMAI, 2016).

Para tanto, apoiou-se em três pontos fundamentais: **transparência**, promovendo o acesso extensivo a "[...] ações, planos, compromissos, decisões, recursos e gastos públicos, entre outros, de maneira tempestiva e em formatos que permitam a população prontamente encontrá-los e (re) utilizá-los" (POSSAMAI, 2016, p. 53-54); **participação**, proporcionando "[...] novos canais e novas oportunidades de envolvimento da população nos assuntos públicos" (*ibid*, p. 54); e **colaboração**, partindo de "[...] espaços de colaboração e inovação junto a cidadãos, empresas, associações, entre outros agentes sociais, para codesenhar [*sic*] e/ou coproduzir [*sic*] soluções para os problemas coletivos, gerando valor público, social e cívico" (*ibid*).

O conceito de Governo Aberto concatena um conjunto de tópicos ligados à "participação cidadã, transparência, colaboração entre governo e sociedade civil, inovações na gestão e na formulação de políticas públicas" (BELIX; GUIMARÃES; MACHADO, 2016, p. 4), de modo a fomentar novas práticas eficientes de gestão, subsidiar o combate à corrupção, fortalecer a democracia e auxiliar o avanço econômico, pontos que o aproximam do ideal Iluminista, no que concerne à ideia de o governo ser colocado à prova, por meio do julgamento da opinião pública (*ibid*). Contudo, para que esse tipo de dado seja aceito pela comunidade, assim como pela *World Wide Web Consortium* (W3C)[6], três pontos devem ser considerados, ou seja, vistos como as leis que o regem: serem encontrados e indexados na *web*, caso contrário, eles não existem; estarem abertos e disponíveis em formato compreensível por máquina, pois, caso contrário, não serão reaproveitados; e serem reproduzíveis, já que, caso um dispositivo legal não permita replicá-los, não são úteis (EAVES, 2009).

[6] Disponível em: https://www.w3.org/.

Hodiernamente, o Governo Aberto transcende a chamada *e-participação* (participação eletrônica) com a disponibilização de ambientes eletrônicos de participação social, como uma espécie de filosofia de integração entre Estado e sociedade, ao passo que viabiliza a reutilização de dados e informações públicos em variadas vias abertas e interoperáveis, o que torna as ações governamentais passíveis de análise e avaliação, para, assim, promover a inovação, entre outras ações (BELIX; GUIMARÃES; MACHADO, 2016; POSSAMAI, 2016).

Ademais, a evolução do conceito de Governo Aberto acompanha uma tendência que vem se fortalecendo desde meados do século XX, isto é, o estabelecimento de movimentos em prol do acesso livre de barreiras às mais diversas informações, como o *Open Access* e as licenças *Creative Commons*[7]. Nessa conjuntura, o entendimento de "aberto", quando se fala de dados no viés governamental em si, vai além da disponibilização de informação e prestação de contas ao cidadão, enquanto ações que integram o conceito de transparência. Em outras palavras, compreende-se a efetiva participação e colaboração do cidadão, mostrando que tal transparência é apenas um dos componentes dessa "abertura", o que infere em medidas participativas e complementares que subsidiam a eficiência na gestão, o fortalecimento da democracia e a inovação (BELIX; GUIMARÃES; MACHADO, 2016; POSSAMAI, 2016; OPEN ARCHIVES INITIATIVE, [2019?]).

Desse modo, países passam a disponibilizar os seus dados, como os EUA, por intermédio da construção de portais de dados abertos governamentais, nacionais e subnacionais, a exemplo do portal *Data.gov*[8], lançado em 2009 com o objetivo de disponibilizar diversos dados do governo, além de conter ferramentas e aplicações para facilitar leituras e análises, podendo ser acessado por qualquer cidadão na Internet (POSSAMAI, 2016).

Consequentemente, o conceito de Governo Aberto expandiu-se pelo mundo, o que não foi diferente no Brasil. No caso do Governo Federal brasileiro, a sua ideia de Governo Aberto é fundamentada em quatro princípios, traduzidos como transparência, prestação de contas, cidadão e tecnologia.

DADOS ABERTOS NO GOVERNO FEDERAL

No Brasil, aliado ao entendimento adotado em países como os EUA, a ideia de Governo Aberto associa-se ao princípio da **transparência**, no qual "as informações sobre as atividades de governo são abertas, compreensíveis, tempestivas, livremente acessíveis e atendem ao padrão básico de dados abertos" (GOVERNO ABERTO, [2020?a]). Outro princípio, chamado de **prestação de contas e responsabilização** (*accountability*), consiste em "[...] regras e mecanismos que estabelecem como os atores justificam suas ações, atuam sobre críticas e exigências e aceitam as responsabilidades que lhes são incumbidas" (*ibid*). No caso do terceiro princípio, da **participação cidadã**, "o governo procura mobilizar a sociedade para debater, colaborar e propor contribuições que levam a um governo mais efetivo e responsivo" (*ibid*). A **tecnologia e inovação**, quarto princípio, transcreve a sentença de que "o governo reconhece a importância das novas tecnologias no fomento à inovação provendo acesso à tecnologia e ampliando a capacidade da sociedade de utilizá-la" (*ibid*).

Em âmbito nacional, tratando-se de dados abertos governamentais, é imperativo citar novamente, a LAI, promulgada em 2011 e destinada à União, aos Estados, aos Municípios e ao Distrito Federal (DF) (BRASIL, 2011; TORINO; TREVISAN; VIDOTTI, 2019), ao apresentar, em seu Art. 3º, as seguintes diretrizes:

> [...] destinam-se a assegurar o direito fundamental de acesso à informação e devem ser executados em conformidade com os princípios básicos da administração pública e com as seguintes diretrizes: I - observância da publicidade como preceito geral e do sigilo como exceção;

[7] Forma de "[...] compartilhamento e uso da criatividade e do conhecimento através de instrumentos jurídicos gratuitos. [...]. As licenças Creative Commons não são contrárias aos direitos de autor. Elas funcionam complementarmente aos direitos autorais e permitem que você modifique seus termos de direitos autorais para melhor atender às suas necessidades" (CREATIVE COMMONS BR, [2020?], destaque nosso).

[8] Disponível em: https://www.data.gov/

> II - divulgação de informações de interesse público, independentemente de solicitações; III - utilização de meios de comunicação viabilizados pela tecnologia da informação; IV - fomento ao desenvolvimento da cultura de transparência na administração pública; V - desenvolvimento do controle social da administração pública (BRASIL, 2011).

Conceitualmente, os dados abertos governamentais assentam-se em oito princípios: **completude**, que estabelece o compromisso de disponibilização total dos dados; **primariedade**, que corresponde à publicação dos dados como foram encontrados em suas fontes de origem; **atualidade**, ou seja, divulgação o mais rápido possível, para preservação da validade; **acessibilidade**, que se trata do dever estar ao alcance do público, independente do propósito; **processabilidade por máquina**, que equivale à garantia de que os dados estejam estrutur ados para que possam ser processados de maneira automatizada; **acessibilidade indiscriminada**, isto é, acesso de todos sem necessidade de identificação ou registro; **formatos não proprietários**, entendidos como a forma para que não haja exclusividade no controle; e **licenciamento livre**, com restrições de privacidade, segurança e controle mínimas, não se prendendo a questões de direitos autorais, marca ou patente (PORTAL BRASILEIRO DE DADOS ABERTOS, [2020?a]).

A OGP consolida-se como "[...] um veículo para se avançar mundialmente no fortalecimento das democracias, na luta contra a corrupção e no fomento a inovações e tecnologias para transformar a governança do século XXI" (GOVERNO ABERTO, [2020?a]). Desde a criação da OGP, o governo brasileiro publicou quatro planos (bienais) de ações. O último (biênio 2018-2020) é composto por 11 compromissos, entre os quais estão o monitoramento e a execução da LAI em Estados e Municípios, e o incremento da participação de segmentos da sociedade no processo legislativo (*ibid*, [2020?b]).

A fim de fundamentar e padronizar o processo de abertura de dados governamentais, estabeleceu-se uma espécie de suplemento à Política de Dados Abertos do Poder Executivo[9], o denominado PDA, o qual deve orientar as organizações ligadas ao Poder Público na implementação do dito processo de abertura de dados, conforme as mais variadas naturezas, como a geoespacial, por exemplo, em prol da disponibilidade, do acesso, da qualidade e reutilização de dados, tanto pela sociedade, quanto pela administração pública (INFRAESTRUTURA NACIONAL DE DADOS ABERTOS, 2020a). Para subsidiar a elaboração do PDA pelas Instituições, foi criada a Resolução nº 3, de 13 de outubro de 2017, a qual determina, a título de ilustração, que sejam considerados o interesse público, as áreas finalísticas institucionais e o desenvolvimento de sistemas de informação de fácil manuseio, com foco no acesso e na rápida compreensão dos dados (BRASIL, 2017).

Considerando a transparência como uma das facetas do Governo Aberto, o governo brasileiro, sob incumbência do Ministério da Transparência e Controladoria Geral da União (CGU), criou, em 2004, o Portal da Transparência do Governo Federal, cujo objetivo é disponibilizar, ao cidadão, dados relativos às contas públicas, isto é, informes sobre como os gastos são empregados, além de informações gerais no tocante à gestão pública. Os dados são enviados periodicamente para a CGU, a depender da natureza de cada um, que os torna acessíveis por meio do portal.

[9] Trata-se de uma política que "[...] define regras para disponibilização de dados abertos governamentais no âmbito do Poder Executivo Federal. Ela é constituída por uma série de documentos normativos, de planejamento e de orientação. Os principais instrumentos que regulam a Política são o Decreto 8.777, de 2016, o Decreto 9.903, de 2019 e a Resolução nº 3 do Comitê Gestor da INDA (CGINDA). O órgão responsável pela gestão e monitoramento da Política é a Controladoria-Geral da União (CGU), por meio da Infraestrutura Nacional de Dados Abertos" (INFRAESTRUTURA NACIONAL DE DADOS ABERTOS, 2020b).

As Instituições responsáveis pelo envio são variadas, como o Sistema Integrado de Administração Financeira do Governo Federal (SIAFI), o Sistema Integrado de Administração de Recursos Humanos (SIAPE), o Banco Central (BC), a Caixa Econômica Federal (CEF), os Ministérios etc. (PORTAL DA TRANSPARÊNCIA, [2020?a]).

As informações contidas no Portal da Transparência[10] concernem ao Poder Executivo, que disponibiliza dados a respeito dos seguintes tópicos: orçamento anual; receitas públicas; despesas públicas; recursos transferidos; gastos por cartão de pagamento; áreas de atuação do governo; programas de governo; benefícios aos cidadãos; programas e ações orçamentárias; emendas parlamentares; órgãos do governo; servidores públicos; viagens a serviço; imóveis funcionais; licitações; contratações; convênios e outros acordos; sanções; relatórios de auditoria; *links* úteis; recursos educativos; e ferramentas interativas para exploração de dados (PORTAL DA TRANSPARÊNCIA, [2020?b]).

Outro exemplo de ambiente digital destinado à divulgação de dados governamentais é o Portal Brasileiro de Dados Abertos[11], que, tal qual o Portal da Transparência, surge como consequência da LAI, que, por sua vez, em seu art. 8º, determina a disponibilização *on-line* de dados em formatos abertos que sejam passíveis de processamento por máquinas (BRASIL, 2011; PORTAL BRASILEIRO DE DADOS ABERTOS, [2020?]). Quanto ao seu escopo, o Portal Brasileiro de Dados Abertos visa à promoção da interlocução entre atores provenientes, tanto da sociedade, quanto da esfera pública, a fim de utilizar dados para o avanço da nação. O conceito de participação social foi levado em conta desde o planejamento até a consecução do sistema, sendo que todas as reuniões foram abertas ao público (*ibid*).

O portal representa um dos primeiros compromissos do Brasil com a OGP, constando, inclusive, no 1º Plano de Ação do Brasil (PAB). Funcionando como um catálogo federado, o portal disponibiliza dados abertos de "[...] saúde suplementar, do sistema de transporte, de segurança pública, indicadores de educação, gastos governamentais, processo eleitoral, etc." (PORTAL BRASILEIRO DE DADOS ABERTOS, [2020?b]). Diversas são as Instituições que contribuem para a alimentação do portal, como o Instituto Brasileiro de Geografia e Estatística (IBGE), a Agência Nacional de Águas (ANA), a Universidade Federal do Piauí (UFPI), a Agência Nacional de Energia Elétrica (ANEEL) etc. (PORTAL BRASILEIRO DE DADOS ABERTOS, [2020?c]).

Não obstante, o cenário dos dados abertos no Governo Federal também se encontra na Justiça Eleitoral, realizada no TSE entre outros órgãos, nesse caso, a partir do uso e da disponibilização de dados sobre candidatos, partidos políticos, eleitores, pleitos etc.

DADOS ABERTOS NA JUSTIÇA ELEITORAL BRASILEIRA

A Justiça Eleitoral é uma justiça especializada que objetiva garantir o respeito à soberania popular e à cidadania, sendo responsável pela organização das eleições e por processos judiciais relacionados ao processo eleitoral (DIAS, 2014). Com base no Código Eleitoral, representado pela Lei nº 4.737, de 1965, em seu art. 12°, o "avaliador A" explica que o TSE, sediado em Brasília, os Tribunais Regionais Eleitorais (TREs), sediados na capital de cada Estado e no DF, as juntas eleitorais e os juízes eleitorais são órgãos da Justiça Eleitoral (BRASIL, 1965).

O processo eleitoral é conduzido da mesma forma em todo o país, com as mesmas regras, mesmo nos níveis estadual ou municipal. As eleições são organizadas em forma de sistema, em que, nesse caso, os órgãos centrais são as unidades do TSE, responsáveis pelas atividades de planejamento, tecnologia da informação (TI), orçamento, recursos humanos, entre outros, sem prejuízo da autonomia de cada TRE (BRASIL, 1994).

[10] Disponível em http://www.portaltransparencia.gov.br/.
[11] Disponível em: http://www.dados.gov.br/.

Ao citar a TI no processo eleitoral, o "servidor A" destaca a informatização das eleições brasileiras, iniciada em 1986 com o recadastramento dos eleitores, passando pela implantação da totalização eletrônica de votos em 1994, e, depois, pela implantação da urna eletrônica em 1996. Na sequência, ocorre a identificação biométrica do eleitor[12], a partir de 2007, que hoje conta com mais de 81,6% dos 146 milhões de eleitores cadastrados biometricamente[13] (TRIBUNAL SUPERIOR ELEITORAL, [2020?a]).

Ao longo da implantação do sistema eletrônico de votação, apuração[14] e totalização, o "servidor A" relata que a Justiça Eleitoral aprimora seus mecanismos de transparência ofertando dados do processo eleitoral para pesquisas e auditorias, ao defender que a transparência é um ponto fundamental, pois suspeitas no pleito podem desestabilizar o país e a democracia. A transparência garante a comprovação da imparcialidade e da legitimidade do processo eleitoral, o que fomenta a disponibilização de dados, encarados como produtos dele, que vão desde o cadastramento e gestão de eleitores, candidatos e partidos, até a coleta, apuração, totalização e divulgação dos votos. Em geral, os dados gerados pelo processo eleitoral são públicos, sendo vedado o acesso aos dados de caráter pessoal do eleitor (TRIBUNAL SUPERIOR ELEITORAL, [2020?a]).

Considerando que as eleições brasileiras são totalmente informatizadas, o "servidor A" assinala que os dados produzidos são trafegados entre sistemas (automatizados), e cada um verifica a integridade das informações recebidas e as autentica para o próximo sistema. Essa automatização retrata as regras do processo eleitoral determinadas pela Constituição Federal, pela legislação e pelas resoluções do TSE, à medida que reduz, ao mínimo necessário, a intervenção humana, tanto para tornar o sistema mais efetivo, quanto para torná-lo mais seguro.

Sobre o sistema eletrônico de votação[15], apuração e totalização, segundo o "servidor A", o recurso envolve um contexto maior que a própria urna eletrônica, pois, considerando a abrangência nacional e a versão única[16] de cada sistema, há uma padronização dos dados gerados em uma eleição, além de resultarem em um grande volume de informação gerada. Consequentemente, em razão da informatização e da necessidade peremptória de transparência, o sistema eletrônico converteu-se em um processo natural de disponibilização de dados eleitorais para o público em geral. Inicialmente, foram disponibilizadas as estatísticas do eleitorado, com informações como sexo, faixa etária e grau de instrução, divididas em vários níveis (país, Unidade da Federação [UF] e município) (TRIBUNAL SUPERIOR ELEITORAL, [2020?b]).

O "servidor A" ainda enfatiza que, diante da crescente demanda de dados solicitados por pesquisadores, sendo, cada demanda, tratada individualmente, respeitando a peculiaridade das solicitações e os diferentes formatos, tipos de dados e abrangências, e da necessidade de os usuários encaminharem *e-mail* ou protocolarem uma solicitação para serem atendidos (TRIBUNAL SUPERIOR ELEITORAL, [2020?b]), o TSE decidiu, então, criar o RDE, para reunir, em um único local, informações diversas sobre o processo eleitoral.

[12] Contempla o cadastramento das impressões digitais e fotografia. Os dados biométricos são utilizados para a verificação biométrica na seção eleitoral e para a identificação biométrica, com o objetivo de eliminar pluralidades e fraudes no cadastro de eleitores.

[13] Segundo o TSE ([2020?a]), o país possui 119.745.607 eleitores cadastrados, do total de 146.758.610 eleitores.

[14] A apuração descreve a soma de votos de uma seção eleitoral.

[15] O "servidor A" afirma que o art. 66, da Lei n ° 9.504, de 1997, descreve, em seu §2º, que os programas serão apresentados em forma de programas-fonte e programas executáveis. Esse dispositivo legal obriga o que se chama de lacração do software, gerando a versão única dos sistemas de preparação e geração de mídias (fora da urna), votação, apuração e sistemas auxiliares (para a urna eletrônica), e totalização (fora da urna eletrônica).

[16] De acordo com o "servidor A", trata-se dos sistemas eleitorais, com algumas poucas exceções, desenvolvidos e padronizados no TSE, sendo que sua utilização se faz por uma única versão, principalmente para os processos de candidaturas, partidos, eleitorado, votação, apuração, totalização e divulgação.

APRESENTAÇÃO DO OBJETO DE ESTUDO: RDE

Ao apresentar o RDE, o "servidor A" informa que a plataforma digital nasce da parceria entre a Secretaria de Tecnologia da Informação (STI), da antiga Seção de Administração de Dados da Coordenadoria de Logística e o Núcleo de Estatística do TSE, da Assessoria de Gestão Estratégica da Diretoria-Geral do TSE. Atualmente, o RDE é gerido pelo referido Núcleo, em conjunto com a Seção de Arquitetura da Informação, da Coordenadoria de Gestão de TI (integrante da STI).

Desde 2009, o RDE disponibiliza dados eleitorais do ano de 1945 em diante para quaisquer cidadãos – como jornalistas, pesquisadores, estudantes e advogados – interessados em dados sobre resultados das eleições, eleitores, partidos políticos, prestações de contas, entre outros. O portal possibilita, por exemplo, a consulta a informações sobre certo candidato, bem como o seu gênero, a sua profissão, escolaridade etc. (TRIBUNAL SUPERIOR ELEITORAL, 2013, [2020?b]).

A periodicidade de atualização é constante, respeitando as retotalizações (TRIBUNAL SUPERIOR ELEITORAL, 2013), em geral, anualmente. Em conformidade com o "servidor A", as informações sobre eleições são geradas a cada pleito ordinário, ou seja, a cada dois anos, alternando-se entre eleições gerais e eleições municipais. Embora os dados de eleitorado sejam disponibilizados anualmente no RDE, é possível consultá-los sumarizados no *website* do TSE (Eleitorado/Atual), a partir de atualização mensal e extraídos do Cadastro Nacional de Eleitores (Cadastro Eleitoral). Outros dados são atualizados conforme sua geração, como é o caso das correspondências esperadas e efetivadas e dos boletins de urna, todos gerados a cada turno eleitoral (TRIBUNAL SUPERIOR ELEITORAL, [2020?b]).

Sobre os dados eleitorais, considerados abertos pelo TSE, o RDE disponibiliza, basicamente, dados de candidatos, eleitores, partidos políticos, pesquisas eleitorais, prestação de contas, processos e resultados dos pleitos.

Nos dados de **candidatos**, de 1945 até 2018, tem-se os nomes e demais dados detalhados, seus bens declarados, a lista de coligações e as vagas para cada cargo eletivo. As informações sobre **comparecimento** e **abstenção** referem-se à participação do eleitorado, nesse caso, do ano de 2018. Quanto ao **eleitorado**, a partir de 1994, seus dados são disponibilizados sumarizados, por município e zona eleitoral, e o seu perfil é apresentado com dados sobre sexo, faixa etária, grau de instrução, indicação de uso de nome social e quantidade de eleitores com deficiência. A esse respeito, nas eleições de 2018, foram disponibilizados dados mais detalhados sobre o tipo de deficiência declarado no cadastramento eleitoral. Os dados de **partidos** (políticos) descrevem todas as siglas partidárias e suas informações, divididas em órgãos e delegados partidários. As **pesquisas eleitorais**, realizadas para um respectivo processo eleitoral, de 2012 até 2020, são disponibilizadas em arquivos no formato *Portable Document Format* (PDF), como cópias das informações encaminhadas pelos institutos de pesquisa. Na **prestação de contas eleitorais**, de 2002 até 2018, constam dados de receitas e despesas de campanha de candidatos, de partidos políticos e de comitês relativas a um pleito eleitoral. A **prestação de contas partidárias**, de 2017 até 2018, apresenta informações da prestação de contas dos partidos políticos, feita anualmente. Os dados do item **processual** concernem ao processo eleitoral, a assuntos específicos, decisões e recursos. Finalmente, os principais dados disponibilizados são os relativos ao processo eletrônico de votação, apuração e totalização, de 1945 até 2018, descritos como **resultados** e distribuídos da seguinte forma: correspondências efetivas e esperadas (primeiro e segundo turnos)[17]; boletim de urna (primeiro e segundo turnos); votação nominal por município e zona; votação em partido por município e zona; votação por seção eleitoral; detalhe da apuração por municípios e zona; e detalhe da apuração por seção eleitoral (TRIBUNAL SUPERIOR ELEITORAL; [2020?b]).

[17] Essas informações permitem, aos partidos políticos e demais interessados, verificar e auditar a origem dos resultados recebidos perante as urnas eletrônicas preparadas em cerimônia pública.

O "servidor A" explica que as **correspondências efetivadas** descrevem as informações entre urna e seção eleitoral efetivamente recebidas. Havendo divergência, o juiz eleitoral registra, no sistema, o motivo da divergência. No caso das **correspondências esperadas**, elas detalham as informações geradas em cada urna eletrônica durante sua preparação. Tais dados associam a identificação única de cada urna a determinadas seção eleitoral, data e hora de preparação e outras informações de segurança. Essas informações sobre cada seção eleitoral são disponibilizadas antes da realização do primeiro ou do segundo turno e utilizadas no sistema de totalização de votos, que, por sua vez, verifica se o boletim de urna recebido pelo sistema se originou da urna eletrônica específica, ou seja, preparada para a sua seção em cerimônia pública. O **boletim de urna**, de sua parte, é constituído dos resultados da apuração de cada seção eleitoral, separados por unidade da federação (UF) e por turno, representando aquilo que foi gerado pela urna eletrônica e transmitido ao sistema de totalização. O "Servidor A" ainda enfatiza que, nas **votações nominais**, em partido e por seção eleitoral, os dados são extraídos do sistema de totalização após o processamento e a contagem dos votos; e, por fim, que a **apuração por município/zona** mostra o perfil da votação conforme o comparecimento e a abstenção dos eleitores, ou seja, a quantidade de votos nominais, brancos e nulos.

No que diz respeito à forma como os dados eleitorais são coletados, o "Servidor A" explica que os sistemas empregados no processo de preparação, votação, apuração e totalização não são as únicas fontes de dados do RDE, pois o Processo Judicial Eletrônico (PJe) provê informações sobre os processos. Os sistemas responsáveis pelas candidaturas, partidos e contas partidárias fornecem as respectivas informações ao RDE, além de outras, advindas de fontes/locais como TSE ou TRE nacionais.

Os dados são coletados por meio de extração semelhante a um *Extract, Transform, Load* (ETL)[18], para que, quase em sua totalidade, as informações sejam transformadas no formato *Comma-Separated Values*[19] (CSV), extensão que se mostrou a mais democrática, pois permite que os interessados façam a carga em banco de dados, ferramentas de tratamento estatístico ou realizem outras formas de transformação.

O entrevistado ("servidor A") esclarece que a disponibilização se dá por dados brutos, os quais podem ser baixados por programas estatísticos, planilhas eletrônicas. Esses dados estão disponibilizados em arquivo compactado (ZIP), que, por sua vez, possui dois arquivos em seu interior, sendo um com os próprios dados eleitorais CSV e outro contendo a descrição dos dados, colocados à disposição no formato PDF. Contudo, recomenda-se que o arquivo PDF "instruções" seja lido observando a data de sua geração, para, assim, realizar importações e consultas corretamente, nesse caso, de responsabilidade do pesquisador (TRIBUNAL SUPERIOR ELEITORAL, 2013, [2020?b]).

Em suma, para o "servidor A", há, no RDE, tanto aspectos do ciclo de vida dos dados, como a coleta e o armazenamento, quanto pilares específicos para estudos sobre dados, como a integridade e a disseminação. Em outras palavras, considerando que as eleições são totalmente automatizadas, todo o fluxo de informações é feito de modo integrado e com verificação de integridade nas trocas e de consistência dos dados armazenados. Isso permite que as informações de transparência sejam extraídas dos sistemas transacionais, transformadas em um formato genérico e armazenadas em repositórios de dados. O RDE representa, então, o conjunto de informações eleitorais disponibilizadas para transparência do processo eleitoral e para pesquisadores interessados no processo.

[18] Do português extração, transformação e carga, "[...] são procedimentos de uma técnica de Data Warehouse (DW), que é responsável pela extracção [sic] de dados de várias fontes, a sua limpeza, optimização e inserção desses dados num DW" (FERREIRA *et al.*, 2010, p. 757).

[19] Valores separados por vírgula.

A própria necessidade de automatização e segurança do processo eletrônico de votação, apuração, totalização, candidaturas e partidos facilita a coerência e integridade dos dados disponibilizados no RDE.

CONCLUSÕES

Iniciativas ocorreram na Internet, promovendo o Movimento de Acesso Aberto, que fomentou acontecimentos como a BOAI[20], ao passo que incentivou a distribuição *on-line* de publicações científicas, na íntegra e de maneira livre. Como resquícios, o supracitado Movimento influenciou na disponibilização dos chamados dados abertos, inclusive, com o mínimo necessário em tecnologia e licenças de uso.

Independentemente dos dados abertos serem públicos ou privados, o importante é que sejam legíveis por máquinas, desconsiderando padrões tecnológicos, econômicos, sociais ou legais. Em outras palavras, os dados abertos devem ser acessados e usados por qualquer sujeito, nesse caso, respeitando fontes originais e licenças. Assim, sem a ambição de controlar e/ou restringir por mecanismos o acesso e uso de dados publicados na Internet, nasce o Governo Aberto, naturalmente exemplificado pela *Transparency and Open Government*, apresentada pelo ex-presidente Obama, e pelo portal *Data.gov*.

O Brasil passa a adotar o Governo aberto, porém respeitando a Constituição Federal Brasileira, nesse caso, no que concerne ao acesso a informações públicas e da maneira mais simples possível, pensamento que se destaca, principalmente, no Portal Brasileiro de Dados Abertos, o qual se baseia na transparência, atitude que fomentou o surgimento do Portal da Transparência, que, por sua vez, surgiu como um dos resultados da LAI.

O RDE, enquanto sistema que disponibiliza dados eleitorais gerados a partir de 1945 para qualquer cidadão, é um exemplo do que se denomina Governo Aberto, ao disponibilizar, sem restrições de acesso, dados sobre eleitores, partidos políticos, candidatos, pesquisas eleitorais, prestação de contas, resultados de pleitos etc., fomentando, assim, a transparência governamental. A atualização dos dados, de ocorrência anual e concernente a um novo pleito ordinário ou sob demanda, à medida que são geradas informações, também constitui um fator que promove a transparência governamental. Outro ponto relevante assenta-se no fato de que o processo eleitoral brasileiro é automatizado e integrado, contando com verificações de integridade, de modo que, quando os dados são extraídos em formatos genéricos (como recomendado pelos princípios associados aos dados abertos governamentais) e depositados no RDE, busca-se a autenticidade, fidedignidade, coerência e clareza de cada segmento informacional relacionado ao processo eleitoral, o que torna o repositório não só importante ferramenta para a transparência da Justiça Eleitoral, mas promotor da democracia brasileira.

Diante do contexto dos dados abertos na esfera governamental, conclui-se que a transparência é o fio condutor do TSE, fomentando o desenvolvimento do RDE, com a finalidade de disponibilizar, com base em tecnologias específicas, dados eleitorais abertos, passíveis de uso em qualquer programa do tipo planilha eletrônica e oriundos do processo eleitoral, desde o cadastramento e a gestão de eleitores até a divulgação dos votos, e, assim, contribuir para a garantia da imparcialidade e legitimidade dos pleitos eleitorais. Por outro lado, este trabalho estimulou o desejo de realização de nova pesquisa, em vista de se investigar, com ênfase, a coleta dos dados pelo RDE.

[20] Iniciativa que "[...] desencadeou uma campanha mundial em prol do acesso aberto (Open Access/OA/AA) a todas as novas publicações científicas revisadas por pares. Esta iniciativa, não criou a ideia do AA. Pelo contrário, procurou deliberadamente reunir projetos já existentes para explorar como poderiam 'trabalhar em conjunto para conseguir o mais amplo, profundo e rápido sucesso'. [...] foi a primeira iniciativa a usar o termo 'open access' para este propósito, a primeira a articular uma definição pública, a primeira a propor estratégias complementares para atingir o AA, a primeira a generalizar o apelo ao AA a todas as disciplinas e países e a primeira a ser acompanhada por financiamento significativo" (BUDAPEST OPEN ACCESS INITIATIVE, [2012?], destaque nosso).

REFERÊNCIAS

BELIX, L.; GUIMARÃES, C. B. S.; MACHADO, J. Qual o conceito de Governo aberto? Uma aproximação aos seus princípios. *In*: CONGRESSO INTERNACIONAL EM GOBIERNO, ADMINISTRACIÓN Y POLÍTICAS PÚBLICAS, 7., 2016, Madrid, *Anais* [...]. Madrid: [*s.n.*], 2016. Disponível em: https://ceweb.br/media/docs/publicacoes/19/Qual%20conceito%20de%20Governo%20Aberto-atualizado_03-out2016.pdf. Acesso em: 25 abr. 2020.

BRANDT, M. B.; VIDOTTI, S. A. B. G.; SANTAREM SEGUNDO, J. E. S. Modelo de dados abertos conectados para informação legislativa. *Informação & Sociedade:* Estudos, João Pessoa, v. 28, n. 2, p. 149-161, maio/ago. 2018. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/37979. Acesso em: 25 abr. 2020.

BRASIL. *Lei n° 4.737, de 15 de julho de 1965*. Institui o Código Eleitoral. Brasília: Presidência da República, 1965. Disponível em: http://www.planalto.gov.br/ccivil_03/LEIS/L4737.htm. Acesso em: 25 abr. 2020.

BRASIL. *Lei n° 8.868, de 14 de abril de 1994*. Regula o acesso a informações previsto no inciso XXXIII do art. 5º , no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. Brasília: Presidência da República, 1994. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 25 abr. 2020.

BRASIL. *Resolução nº 3, de 13 de outubro de 2017*. Aprova as normas sobre elaboração e publicação de Planos de Dados Abertos, conforme disposto no Decreto nº 8.777, de 11 de maio de 2016. [Brasília: Ministério do Planejamento, Desenvolvimento e Gestão], 2017. Disponível em: http://wiki.dados.gov.br/GetFile.aspx?File=%2fComiteGestor%2fResolu%C3%A7%C3%B5es%2fresolucao-cginda-3-13-10-2017.pdf. Acesso em: 25 abr. 2020.

BUDAPEST OPEN ACCESS INITIATIVE. *Dez anos da Iniciativa de Budapeste em Acesso Aberto*: a abertura como caminho a seguir. Tradução de Carolina Rossini. [Hungria: *s.n.*], [2012?]. Disponível em: https://www.budapestopenaccessinitiative.org/boai-10-translations/portuguese-brazilian-translation. Acesso em: 13 set. 2020.

CARDOSO, P. H. *Ciência de dados aplicada a dados governamentais abertos sob a ótica da Ciência da Informação*. 2019. 110 f. Dissertação (Mestrando em Ciência da Informação) - Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2019.

CREATIVE COMMONS BR. *Sobre*. [*S. l.: s. n.*], [2020?]. Disponível em: https://br.creativecommons.org/sobre/. Acesso em: 18 mar. 2019.

DIAS, R. L. A. de B. Justiça Eleitoral: composição, competências e funções. *Revista Eletrônica da Escola Judiciária Eleitoral*, Brasília, DF, n. 1, ano 4, 2014. Disponível em: http://www.tse.jus.br/o-tse/escola-judiciaria-eleitoral/publicacoes/revistas-da-eje/artigos/revista-eletronica-eje-n.-6-ano-2. Acesso em: 20 mar. 2020.

EAVES, D. *The Three Laws of Open Government Data*. [*S. l.*]: eaves.ca, set. 2009. Disponível em: https://eaves.ca/2009/09/30/three-law-of-open-government-data/. Acesso em: 27 abr. 2020.

FERREIRA, J. *et al*. O Processo ETL em Sistemas Data Warehouse. *In*: INFORUM SIMPÓSIO DE INFORMÁTICA, 2010, 2., 2010, Braga, PT, [=*Anais* **[...]** Braga, PT: 2010. Disponível em: https://www.researchgate.net/publication/265195317_O_Processo_ETL_em_Sistemas_Data_Warehouse. Acesso em: 20 mar. 2020.

FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO. *Metabuscador de dados de pesquisa.* [São Paulo: *s. n.*], 2019. Disponível em: https://metabuscador.uspdigital.usp.br/. Acesso em: 26 abr. 2020.

GOV.BR. *Dados Abertos*. [*S. l.*]: Governo Digital, [2020?]. Disponível em: https://www.gov.br/governodigital/pt-br/dados-abertos/. Acesso em: 27 abr. 2020.

GOVERNO ABERTO. *4º Plano de ação brasileiro*. [*S. l.*: *s. n.*], [2020?b]. Disponível em: http://governoaberto.cgu.gov.br/a-ogp/planos-de-acao/4o-plano-de-acao-brasileiro. Acesso em: 25 abr. 2020.

GOVERNO ABERTO. *O que é governo aberto*. [*S. l.*: *s. n.*], [2020?a]. Disponível em: https://governoaberto.cgu.gov.br/governo-aberto-no-brasil/o-que-e-governo-aberto. Acesso em: 25 abr. 2020.

INFRAESTRUTURA NACIONAL DE DADOS ABERTOS. *Plano de Dados Abertos*. Brasília, DF: Ministério do Planejamento, Desenvolvimento e Gestão, 2020a. Disponível em: http://wiki.dados.gov.br/Default.aspx?Page=Plano-de-Dados-Abertos&NavPath=Principais%20t%C3%B3picos. Acesso em: 25 abr. 2020.

INFRAESTRUTURA NACIONAL DE DADOS ABERTOS. *Política de Dados Abertos do Poder Executivo Federal.* Brasília, DF: Ministério do Planejamento, Desenvolvimento e Gestão, 2020b. Disponível em: http://wiki.dados.gov.br/Politica-de-Dados-Abertos.ashx. Acesso em: 25 abr. 2020.

ISOTANI, S.; BITTENCOURT, I. I. *Dados Abertos Conectados*. São Paulo: Novatec Editora, 2015.

KAPLAN, C. A. *A guide to transparency in election administration*. [*S. l.*]: IFES Consulation, [2006]. Disponível em: https://aceproject.org/ero-en/topics/voter-registration/vrx_o005.pdf. Acesso em: 29 abr. 2020.

OPEN ARCHIVES INITIATIVE. *About OAI.* [*S. l.*]: OAI, [2019?]. Disponível em: https://www.openarchives.org/organization/. Acesso em: 11 dez. 2019.

OPEN KNOWLEDGE INTERNATIONAL. O que são dados abertos? *In:* OPEN KNOWLEDGE INTERNATIONAL. *Open Data HandBook*. [*S. l.: Open Knowledge International*], [2020?]. Disponível em: http://opendatahandbook.org/guide/pt_BR/what-is-open-data/. Acesso em: 23 abr. 2020.

PORTAL BRASILEIRO DE DADOS ABERTOS. *Biometria*. [Brasília, DF: TSE], [2020?a]. Disponível em: http://www.justicaeleitoral.jus.br/biometria/. Acesso em: 29 abr. 2020.

PORTAL BRASILEIRO DE DADOS ABERTOS. *Busque no Portal*. Brasília, DF: Ministério do Planejamento, Desenvolvimento e Gestão, [2020?c]. Disponível em: http://www.dados.gov.br/. Acesso em: 25 abr. 2020.

PORTAL BRASILEIRO DE DADOS ABERTOS. *O que são dados abertos?* Brasília, DF: Ministério do Planejamento, Desenvolvimento e Gestão, [2020?a]. Disponível em: http://www.dados.gov.br/pagina/dados-abertos . Acesso em: 25 abr. 2020.

PORTAL BRASILEIRO DE DADOS ABERTOS. *O que você encontra no Portal*. [Brasília, DF]: CGU, [2020?b]. Disponível em: http://www.portaltransparencia.gov.br/sobre/o-que-voce-encontra-no-portal. Acesso em: 25 abr. 2020.

PORTAL BRASILEIRO DE DADOS ABERTOS. *Relato da iniciativa*. [*S. l.: s. n.*], [2020?]. Disponível em: https://www.gov.br/governodigital/pt-br/dados-abertos/portalbrasileirodadosabertos.pdf. Acesso em: 25 abr. 2020.

PORTAL BRASILEIRO DE DADOS ABERTOS. *Repositório de Dados Eleitorais*. [Brasília, DF: TSE], [2020?b]. Disponível em: http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1/repositorio-de-dados-eleitorais. Acesso em: 26 abr. 2020.

PORTAL BRASILEIRO DE DADOS ABERTOS. *Sobre o dados.gov.br.* Brasília, DF: Ministério do Planejamento, Desenvolvimento e Gestão, [2020?b]. Disponível em: http://www.dados.gov.br/pagina/sobre . Acesso em: 25 abr. 2020.

PORTAL DA TRANSPARÊNCIA. *O que é e como funciona*. [Brasília, DF]: CGU, [2020?a]. Disponível em: http://www.portaltransparencia.gov.br/sobre/o-que-e-e-como-funciona. Acesso em: 25 abr. 2020.

POSSAMAI, A. J. *Dados abertos no governo federal brasileiro*: desafios de transparência e interoperabilidade. 2016. 300 f. Tese (Doutorado em Ciência Política) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016. Disponível em: https://www.lume.ufrgs.br/handle/10183/156363. Acesso em: 25 abr. 2020.

SANTAREM SEGUNDO, J. E. Tecnologías de la información y la comunicación para proporcionar datos abiertos en formato semántico. *Ibersid***:** revista de sistemas de informacipon y documentación, Zaragoza, v. 7, p.33-40, 2013. Disponível em: https://www.ibersid.eu/ojs/index.php/ibersid/article/view/4075. Acesso em: 24 abr. 2020.

THE OPEN DEFINITION. *Open Definition 2.1*. [*S. l.: s. n.*], [2020?]. Disponível em: http://opendefinition.org/. Acesso em: 23 abr. 2020.

TORINO, E.; TREVISAN, G. L.; VIDOTTI, S. A. B. G. Dados abertos CAPES: um olhar à luz dos desafios para publicação de dados na web. *Ciência da Informação*, Brasília, DF, v. 48, n. 3, p. 38-46, set./dez. 2019. Disponível em: http://revista.ibict.br/ciinf/article/view/4866/4428. Acesso em: 25 abr. 2020.

TRIBUNAL SUPERIOR ELEITORAL. Assessoria de Comunicação. Banco de dados do TSE disponibiliza consulta a dados de eleições realizadas a partir de 1950. *[Boletim TSE]*, Brasília, DF, 16 ago. 2013. Disponível em: http://www.tse.jus.br/imprensa/noticias-tse/boletim/banco-de-dados-do-tse-disponibiliza-consulta-a-dados-de-eleicoes-realizadas-a-partir-de-1950. Acesso em: 26 abr. 2020.

AGRADECIMENTOS

Ao Núcleo de Estatística da Assessoria de Gestão Estratégica do TSE, por sanar, prontamente, dúvidas sobre o RDE.

Modelo de Análise Temporal em Contexto Semântico de Gerenciamento de Emergências

Gustavo Marttos Cáceres Pereira
Mestre em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp) - Brasil. Engenheiro de Dados da Will Bank (WILL BANK) - Brasil.
http://lattes.cnpq.br/4119044801375917
E-mail: gustavo.marttos@unesp.br

Leonardo Castro Botega
Pós-Doutorado pelo Instituto de Ciências Matemáticas e de Computação (ICMC-USP) - Brasil. Doutor em Ciência da Computação pela Universidade Federal de São Carlos (UFSCAR) – SP - Brasil. Professor da Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp) - Brasil.
http://lattes.cnpq.br/6027755717265622
https://orcid.org/0000-0003-1495-5935
E-mail: leonardo.botega@unesp.br

Submetido em: 25/11/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021 .

RESUMO

No contexto de gerenciamento de emergências, no qual as informações são provenientes de fontes heterogêneas, é necessário que as tomadas de decisões sejam assertivas e dentro de um intervalo de tempo hábil. O tempo possui grande relevância, visto ser fundamental no domínio desse contexto, uma vez que é criando uma linha do tempo, ou seja, no decorrer dele, que se torna viável a percepção e compreensão de todas as características de uma situação. A qualidade da informação torna-se imprescindível no contexto de gerenciamento de emergências, considerando a complexidade e dinamicidade dos dados. Sendo assim, este trabalho visa à melhoria dos processos informacionais da recuperação da informação, por meio da criação de um modelo de análise temporal no contexto semântico. A natureza deste trabalho é qualitativa, de finalidade teórico-aplicada e de tipo exploratória. A sua metodologia envolve situações de emergência de incêndio, em que foi possível avaliar o comportamento da qualidade de informação e inferir novos conhecimentos temporais que poderão servir de insumos para tomadas de decisões mais categóricas.

Palavras-chave: Qualidade da informação. Ontologia. Gerenciamento de rmergências. Recuperação da informação.

Time Analysis Model in Semantic Context of Emergency Management

ABSTRACT

In the context of emergency management, where information comes from heterogeneous, complex and dynamic sources, decision making is required to be assertive and within a timely interval of time. Time has great relevance, being fundamental in the domain of this context, because it is creating a timeline, that is to say during it, the perception and comprehension of all the characteristics of a situation becomes viable. The quality of information becomes indispensable in the context of emergency management, mainly by dynamic and complex factors. This work aims to improve the informational processes of information retrieval through the creation of a model of temporal analysis in the semantic context. The nature of this work is qualitative, of theoretical-applied purpose and exploratory, and its methodology involving fire emergency situations, in which it was possible to evaluate the behavior of information quality and infer new temporal knowledge that may serve as inputs for making more assertive decisions.

Keywords*: Information quality. Ontology. Emergency management. Information retrieval.*

Modelo de Análisis de Tiempo en el Contexto Semántico de la Gestión de Emergencias

RESUMEN

En el contexto de la gestión de emergencias, donde la información proviene de fuentes heterogéneas, es necesario que la toma de decisiones sea asertiva y dentro de un intervalo oportuno. El tiempo tiene gran relevancia porque es fundamental en el ámbito de este contexto, porque está creando una línea de tiempo, es decir, en el curso de esta, que se hace factible percibir y comprender todas las características de una situación. La calidad de la información se vuelve esencial en el contexto de la gestión de emergencias, teniendo en cuenta la complejidad y dinámica de los datos. Este trabajo tiene como objetivo mejorar los procesos informativos de recuperación de información a través de la creación de un modelo de análisis temporal en el contexto semántico. La naturaleza de este trabajo es cualitativa, de propósito teórico-aplicado y exploratorio, y su metodología que implica situaciones de emergencia contra incendios, en la que fue posible evaluar el comportamiento de la calidad de la información e inferir nuevos conocimientos temporales que pueden servir como insumos para tomar decisiones más asertivas.

Palabras clave*: Calidad de la información. Ontologías. Gestión de emergencias. Recuperación de información.*

INTRODUÇÃO

Consciência Situacional (do inglês *Situational Awareness* - SAW) refere-se a um estado cognitivo importante para auxiliar tomadores de decisão que atuam em ambientes informacionais complexos e dinâmicos em uma variedade de domínios, entre eles, o de gerenciamento de emergências.

Especificamente, o conceito de SAW concerne à percepção dos elementos em um ambiente analógico ou digital diante de um volume de espaço e tempo; à compreensão do significado, da relevância e importância de seus aspectos individuais e coletivos; e à projeção de seus status em um futuro próximo (KOKAR; ENDSLEY, 2012).

Indivíduos como os operadores de sistemas de gerenciamento de emergências, os gestores de unidades de conservação e os bombeiros estão constantemente sob alta pressão e expostos a uma gama de informações sensíveis. Logo, necessitam manter seus níveis de SAW elevados para assim sustentar o melhor retrato de uma situação crítica e tomar a decisão mais assertiva, evitando prejuízos à vida, ao patrimônio e ao meio ambiente. SAW não garante que a melhor decisão será tomada, entretanto, garante melhores subsídios para que possa melhorá-la.

Para suportar a obtenção e manutenção de SAW, foi constatado que o uso de modelos semânticos, mais especificamente ontologias, quando aplicados para suportar sistemas de gerenciamento de emergências, podem contribuir para uma melhor assertividade nas inferências úteis à tomada de decisão (MATHEUS; KOKAR; BACLAWSKI, 2003).

Ademais, as ontologias favorecem buscas por significados ao invés de termos literais, como ocorre em modelos sintáticos. Inferir novas informações utilizando axiomas e regras estabelecidas, além da contextualização já presente na ontologia, possibilita que as necessidades informacionais de um indivíduo sejam supridas.

Na Ciência da Informação, o impacto de suprir as necessidades informacionais já é algo pertinente à Recuperação da Informação, o que permite estender o olhar sob o prisma da interação e integração entre as fontes informacionais e quem as consomem.

Ambientes informacionais no contexto de gerenciamento de emergências possuem dados e informações complexas, dinâmicas e são provenientes de fontes heterogêneas. Tais características são decorrentes dos problemas da qualidade de dados e informação, as quais, ao serem recuperadas por processos de recuperação da informação, são prejudiciais para o estímulo da SAW de um indivíduo, visto que este pode não ter informações precisas, atualizadas, completas e consistentes.

Além disso, há a necessidade do envolvimento da dimensão temporal, sendo um artefato de grande importância e relevância para o domínio supracitado, bem como para a Ciência da Informação. A atualidade, enquanto dimensão de qualidade da informação, é uma característica intrinsecamente relacionada ao tempo, afinal, as informações mudam no decorrer do tempo e, portanto, criam uma linha temporal que pode ser analisada para viabilizar a recuperação da informação, proporcionando melhores insumos informacionais para uma tomada de decisão mais categórica.

Ainda que a qualidade esteja presente, ela não viabiliza a contextualização, a descoberta de novas informações, tampouco a viabilidade de se criar uma linha do tempo de acordo com a semântica informacional. Outrossim, a combinação de semântica e temporalidade para a recuperação da informação favorece o entendimento de situações críticas e a respectiva tomada de decisão feita por humanos operadores de emergências, o que, consequentemente, também favorece os níveis de SAW pertencentes aos operadores.

A complexidade e a dinâmica de um ambiente informacional de avaliação de emergências aumentam proporcionalmente, conforme a quantidade de variáveis existentes nesse ambiente, tornando a aquisição e manutenção de SAW processos mais difíceis de serem atingidos.

Ademais, as informações desses ambientes tendem a ter problemas de qualidade, prejudicando o consumo informacional para os tomadores de decisão. Partindo dessa premissa, foi identificado que a qualidade da informação e a temporalidade formam um conjunto coerente e suscetível à assistência dada à recuperação da informação, proporcionando melhores insumos, a fim de suprir as necessidades informacionais daqueles que os utilizarem.

Diante da contextualização e dos desafios retratados, a Ciência da Informação torna-se fundamental perante a comunicação, interação e integração de saberes proporcionados por disciplinas de outras áreas. Esse elo, segundo Saracevic (1996), refere-se à interdisciplinaridade natural da Ciência da Informação.

Na literatura, é possível identificar a presença de trabalhos que lidam com ambientes informacionais no contexto de gerenciamento de emergências junto à interdisciplinaridade e a propostas voltadas à Ciência da Informação (MELO; BOTEGA; SANTARÉM SEGUNDO, 2017; OLIVEIRA *et al.*, 2017; SILVA *et al.*, 2018; BOTEGA *et al.*, 2019; PEREIRA JUNIOR; PEREIRA; BOTEGA, 2019).

Diante disso, este trabalho discorre a respeito de um modelo de especificação semântica e temporal, buscando observar como a qualidade da informação e o tempo podem influenciar os resultados dos insumos informacionais a serem consumidos por indivíduos, para, desse modo, elevar continuamente sua SAW, sustentando o melhor retrato de uma situação crítica para tomar a decisão mais assertiva.

Este trabalho tem, como objetivo geral, a criação de um modelo que possibilite descobertas informacionais baseadas em tempo, a partir de inferências e regras ontológicas, e que também permita a melhoria dos processos informacionais da recuperação da informação. Especificamente, busca-se definir a estrutura do processo informacional que contemple a qualidade da informação, a semântica e a temporalidade, bem como incluir a dimensão temporal e reutilizar vocabulários específicos em uma ontologia de domínio ciente da qualidade desenvolvida em trabalhos anteriores.

O arcabouço metodológico se sustenta em virtude da natureza qualitativa, de finalidade teórico-aplicada e de tipo exploratória. A pesquisa é orientada por situações de emergência de incêndio instanciadas em uma ontologia de domínio no contexto de gerenciamento de emergências, onde foi realizada a avaliação da qualidade da informação e a dimensão temporal foi incluída em sua estrutura.As ontologias na recuperação da informação serão apresentadas na Seção 2; as ontologias temporais, na Seção 3; as contextualizações de qualidade da informação, na Seção 4; a metodologia para avaliação da qualidade de dados informações, na Seção 5; o modelo de análise temporal em contexto semântico de gerenciamento de emergências, na Seção 6; a prova de conceito, na Seção 7; e, por fim, as considerações finais, na Seção 8.

ONTOLOGIAS NA RECUPERAÇÃO DA INFORMAÇÃO

A recuperação da informação visa a buscar informações em um *corpus* de documentos, por meio de expressões de consultas formatadas com um conjunto de palavras-chave. Entretanto, um dos problemas que percorrem a área é conseguir distinguir o que é relevante e irrelevante. De acordo com Paz-Trillo, Wassermann e Braga (2005), apesar do indivíduo ter a capacidade de informar o que pode ser relevante ou não, os mecanismos de um sistema de recuperação da informação tendem a não conseguir atingir um grau significante de precisão, como visto nos modelos apresentados anteriormente.

Para os autores, a solução é a construção de uma estrutura de conceitos ao invés do uso de palavras-chave, pois estas se referem a sinônimos, os quais são polissêmicos, ou seja, podem expressar um conceito diferente, enquanto se espera outro conceito a ser retornado. Essa estrutura de conceitos diz respeito às ontologias computacionais, neste trabalho, consideradas apenas ontologias, as quais Gómez-Pérez e Benjamins (1999) conceituam como um conjunto de termos hierarquicamente ordenados para descrever um domínio que possa ser utilizado como princípio para uma base de conhecimento.

O uso de ontologias possibilita uma gama de opções no quesito da recuperação da informação, a exemplo da inferência de novos conhecimentos por intermédio de axiomas e regras estabelecidas, bem como da contextualização das necessidades informacionais do indivíduo e seus respectivos significados.

A proposta do uso de ontologias como modelos de recuperação da informação se dá pela busca por significados ao invés de termos literais, como ocorre nos modelos clássicos.

Ademais, para Wiegand e García (2007), as ontologias auxiliam no processo da interoperabilidade semântica, ou seja, todos os sistemas que utilizam a ontologia podem estar em conformidade com metadados adotados por ela. Caso não estejam, a ontologia servirá como autoridade para definir quais metadados serão adotados no que concerne aos vínculos necessários para o uso no sistema. Isso permite que o indivíduo utilize diversos metadados para construir sua expressão de busca. Entretanto, estes devem ser definidos de maneira apropriada a fim de que não dificultem o processo de recuperação da informação.

Os autores ainda debatem sobre outros aspectos, que tornam viável o uso de ontologias na recuperação da informação, como os modelos clássicos, que possuem demasiados filtros e, às vezes, existem palavras-chave na expressão de busca que não têm sentido perante o contexto dos documentos do *corpus*. Além disso, quando há muitos documentos como resultado, isso pode atrapalhar o indivíduo, pois este acaba por confundir suas necessidades informacionais (WIEGAND; GARCÍA, 2007).

ONTOLOGIAS TEMPORAIS

Tao *et al*. (2010) discorrem sobre a importância da inclusão da dimensão temporal em uma ontologia que já possui alguma característica qualitativa. O ponto de vista dos autores, essa dimensão é fundamental para o raciocínio temporal, isto é, corresponde a respostas que podem mudar no decorrer do tempo e criam, portanto, uma linha do tempo que pode ser analisada durante sua recuperação e representação, e, consequentemente, pode servir de insumo para que os operadores tenham uma melhor percepção e compreensão das informações.

Não obstante, Okeyo, Chen e Wang (2014) reiteram a importância do relacionamento temporal, pois, de acordo com seus estudos, representar conhecimento temporal usando OWL é um desafio, haja vista que essa tecnologia suporta apenas relações unárias e binárias, enquanto uma relação temporal depende de, no mínimo, uma relação ternária.

Com o relacionamento temporal estabelecido, pode-se inferir novas informações, obtendo conhecimento temporal. Para tanto, é necessário que todas as instâncias estejam com seus atributos granulares, porque assim a linha do tempo pode se formar.

Uma granularidade é a normalização de datas, ou seja, é deixar as datas de modo que sejam interpretáveis por mecanismos computacionais. Uma expressão de tempo dada por "dois dias atrás" deve ser normalizada para "2020-03-12", caso o dia corrente seja "2020-03-14", por exemplo. Outras expressões, como "antes", "depois" e "durante", também são válidas (HASANUZZAMAN *et al*., 2014).

Segundo Tao *et al*. (2010), a dimensão temporal em relação à análise de dados emergenciais possui diversas aplicabilidades, tais como: (1) a descoberta de padrões temporais em uma situação de incêndio florestal em determinado bioma; (2) a explicação de situações passadas, buscando trazer as prováveis causas que levam a situações de emergência; e (3) a projeção de estados futuros, como a possibilidade do fogo de um incêndio florestal se alastrar para outras áreas.

QUALIDADE DA INFORMAÇÃO

Devido à demasiada quantia de dados e informações presentes nos ambientes informacionais, principalmente os contextualizados no gerenciamento de emergências, a qualidade da informação torna-se imprescindível para que ela seja avaliada conforme as necessidades informacionais dos indivíduos.

A definição de qualidade é subjetiva, em outros termos, é necessário alinhar as necessidades informacionais, as ações e os objetivos de cada domínio a que a qualidade está contextualizada. Na visão de Buckland (1991), que enfatiza a informação como coisa, tais características pertinentes à qualidade, ao serem consideradas livres de quaisquer problemas e/ou falhas, passam a ser válidas perante a avaliação de informação, a qual utiliza a qualidade como critério (OLETO, 2006; PEREIRA JUNIOR; PEREIRA; BOTEGA, 2019).

Para ser considerada bem-sucedida em suas aplicações, a qualidade depende de que o indivíduo esteja com suas necessidades informacionais alinhadas ao domínio em que está situado, isto é, os critérios utilizados pelo indivíduo para avaliar a qualidade da informação devem estar presentes no domínio. Olson (2003) argumenta que a informação é de qualidade se satisfizer os requisitos informacionais para seu uso; logo, se não satisfaz, requer uma melhor qualidade.

Nehmy e Paim (1998) e Oleto (2006) entendem a qualidade da informação como um conjunto de dimensões relacionadas, mensuráveis e multidimensionais, podendo existir diversas relações entre elas, tais como abrangência, acessibilidade, atualidade, objetividade, precisão e validade. Calazans (2008), por sua vez, discorre a respeito da ausência da qualidade da informação, a qual pode causar impactos, afetando diretamente no uso da informação, exigindo, portanto, que se providencie soluções o quanto antes.

Na concepção de Oliveira *et al*. (2017) e Silva *et al*. (2018), ambientes informacionais presentes no domínio de gerenciamento de emergências lidam com informações complexas, heterogêneas, imprevisíveis e dinâmicas, e, por isso, limitam a representabilidade e recuperabilidade da informação diante da qualidade, justamente pelo fato das informações estarem incompletas, imprecisas e difusas.

Botega *et al*. (2019) complementam que a qualidade da informação pode beneficiar tanto os processos automatizados, quanto a compreensão humana perante a situações de emergência. A presença das dimensões qualitativas pode auxiliar os indivíduos envolvidos no que diz respeito à confiabilidade informacional.

Concernente às dimensões e métricas qualitativas, é válido ressaltar que as dimensões qualitativas mencionadas possuem descrições e definições similares, porém a aplicabilidade e o funcionamento de cada uma variam de acordo com o domínio em que estão sendo aplicadas. Cada dimensão representa um problema de qualidade em relação à sua aplicação, como, por exemplo, as já mencionadas atualidade, precisão, completude e acessibilidade (MELO; BOTEGA; SANTARÉM SEGUNDO, 2017).

Botega (2016) define dimensão como um artefato composto por objetivos, tarefas e decisões associadas, segundo os requisitos e categorizações de qualidade presentes em cada área de aplicação. Isto é, as dimensões tornam-se precisas, identificáveis, mensuráveis e quantificáveis.

Consoante Liu e Chi (2002), de uma perspectiva teórico-específica, as dimensões podem ser utilizadas a partir de três abordagens: (1) intuitiva, sendo baseada na experiência de especialistas de um domínio; (2) empírica, a qual tem seus atributos determinados pelos indivíduos presentes em um domínio; e (3) teórica, à medida que enfatiza teorias previamente estabelecidas e pesquisas operacionais.

A última abordagem proposta por Liu e Chi (2002) sintetiza a informação como coisa ou produto, em consonância com os trabalhos de Buckland (1991), Nehmy e Paim (1998) e Oleto (2006). Sendo assim, neste trabalho, serão adotadas as abordagens intuitiva e teórica, pois ambientes informacionais situados em domínio críticos (como o gerenciamento de emergências) demandam a existência de especialistas para as respectivas definições das dimensões qualitativas, bem como para o embasamento em estudos prévios acerca desse mesmo domínio.

As métricas qualitativas, por sua vez, referem-se à forma como as dimensões serão mensuradas, partindo da premissa de que sua aplicabilidade dependerá do contexto e do domínio em que a informação a ser avaliada está inserida.

METODOLOGIA PARA AVALIAÇÃO DA QUALIDADE DE DADOS E INFORMAÇÕES

Devido ao montante de dados e informações, tornou-se necessário o uso de metodologias de avaliação e gestão de qualidade. Batini *et al*. (2009) apresentam um conjunto de metodologias que servem para diversos domínios.

Neste trabalho, será adotada a metodologia IQESA (*Information Quality Assessment Methodology in the Context of Emergency Situational Awareness*), proposta por Botega *et al*. (2017) para a avaliação de qualidade de dados e informações provenientes de situações de emergências.

Os requisitos adotados por essa metodologia para a concepção de qualidade são especificados por especialistas no domínio em que a mesma será aplicada, conforme a abordagem intuitiva de Liu e Chi (2002).

A IQESA, enquanto instrumento para avaliação da qualidade, permite a ilustração de todas as fases para avaliá-la e representá-la como parte de um processo de avaliação de informações no contexto de gerenciamento de emergências. Em relação às dimensões, cada dimensão conta com uma expressão matemática para auxiliar em sua quantificação, utilizando suas respectivas métricas.

A IQESA viabiliza o monitoramento das mudanças na qualidade da informação por meio de métodos de fusão de informações. Essa metodologia também se propõe a ser flexível, acompanhando os sistemas que a implementam por intermédio das atualizações que delas decorrem.

A metodologia conta com três etapas básicas: (1) elucidação dos requisitos da qualidade de dados e informações; (2) definição de funções e métricas para quantificar as dimensões; e (3) representação da informação situacional qualificada.

Como exemplo prático, Silva *et al*. (2018) propuseram uma ontologia ciente de qualidade para o domínio de gerenciamento de emergências. A metodologia de avaliação e representação de qualidade adotada foi a IQESA, que possibilitou aos autores utilizarem as dimensões atualidade, completude, consistência, relevância e certeza.

A avaliação da qualidade da informação pode influenciar diretamente no processo de aquisição e manutenção de SAW de um indivíduo e, consequentemente, na tomada de decisão em si, a qual necessita de insumos informacionais precisos, atuais e completos para que seja mais assertiva.

Entretanto, sabe-se que a origem de tais insumos informacionais não possui índices qualitativos aceitáveis perante às recomendações de um especialista do domínio de gerenciamento de emergências, pois as fontes informacionais são heterogêneas e complexas, tornando a qualidade da informação, em sua origem, precária e, somada aos fatores de estresse e modelos mentais dos indivíduos em uma situação de emergência, inviável de ser levada em consideração, o que pode gerar riscos à vida, ao patrimônio e ao meio ambiente.

Atrelar a qualidade da informação torna-se interessante, na medida em que pode viabilizar insumos informacionais mais apropriados e que melhor atendam às necessidades informacionais dos indivíduos.

Conforme apresentado na seção anterior, é possível e viável o uso de ontologias como modelo semântico para a recuperação da informação. Melhorar a qualidade de informações provenientes de uma ontologia, que, em sua essência, possibilita a descoberta de novas informações por meio de inferências e regras definidas em axiomas, fornece critérios e parâmetros baseados em índices qualitativos, uma vez que a qualidade reflete o valor e o quão confiável, atual, precisa e compreensível é uma informação.

Todas as dimensões e métricas aplicadas remetem à percepção humana daquilo que se espera da informação. Ela advir de uma ontologia ciente de qualidade é instigante, pois a contextualização informacional oriunda da ontologia somada aos fatores qualitativos pode resultar em melhores insumos para que o indivíduo melhore continuamente sua SAW.

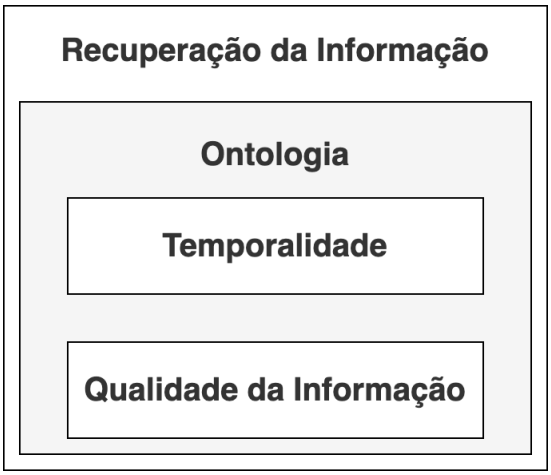
Silva *et al*. (2018) desenvolveram uma ontologia que possui meios para avaliar a qualidade da informação, objetivando produzir melhores insumos informacionais para o desenvolvimento de SAW. Os autores elaboraram um estudo de caso envolvendo um atendimento a uma situação de incêndio florestal, que foi utilizado para demonstrar a aplicabilidade da ontologia junto à gestão da qualidade, contribuindo positivamente para uma SAW dos operadores humanos envolvidos.

MODELO DE ANÁLISE TEMPORAL EM CONTEXTO SEMÂNTICO DE GERENCIAMENTO DE EMERGÊNCIAS

Esta seção visa a apresentar um modelo de análise temporal em contexto semântico de gerenciamento de emergências, objetivando melhorar os processos informacionais da recuperação da informação, uma vez que possibilitam a descoberta de novas informações por meio de inferências e definições de axiomas bem estruturados. Ademais, a qualidade da informação, em conjunto com a dimensão temporal, favorece a atualidade, a relevância e a credibilidade informacionais, podendo melhorar os documentos recuperados pela busca feita por um indivíduo, o que, consequentemente, atenderá a suas necessidades informacionais.

A figura 1 contempla a explanação acima, tendo a temporalidade e a qualidade da informação englobadas pela ontologia, enquanto esta serve como fonte informacional para a recuperação da informação.

Figura 1 − Modelo de Análise Temporal em Contexto Semântico de Gerenciamento de Emergências



Fonte: Autores.

A Recuperação da Informação depende de um modelo que permita atender às necessidades de indivíduos situados nesse domínio. Portanto, a ontologia de domínio, enquanto modelo semântico, favorece diretamente a contextualização de dados, tornando, assim, os processos informacionais da Recuperação da Informação mais eficientes. A interação e integração entre ambos os elementos do modelo possibilitam que as informações recuperadas transportem consigo toda a semântica que a ontologia proporciona, possibilitando o enriquecimento do conhecimento do indivíduo, caso este tenha suas necessidades informacionais supridas.

A ontologia de domínio, entretanto, demanda que os dados tenham seus índices de qualidade fidedignos perante a necessidade informacional de um indivíduo. Logo, a Qualidade da Informação se faz presente para atender à percepção humana do que espera do insumo informacional. Por sua vez, o espectro da temporalidade, unido à Qualidade da Informação, permite a estruturação de uma linha temporal, tornando análises informacionais mais relevantes, uma vez que a dimensão temporal representa um artefato de grande importância e relevância no domínio de gerenciamento de emergências.

A metodologia aqui empregada para avaliação da qualidade no modelo é a IQESA, proposta por Botega *et al.* (2017). Esta objetiva avaliar e representar os índices de dados e qualidade da informação do melhor modo, possibilitando o uso da qualidade em um processo de avaliação da informação.

O processo da IQESA é organizado e dividido em três etapas: (1) levantamento de requisitos de qualidade dos dados e das informações; (2) definição das funções e métricas para quantificar as dimensões de qualidade; e (3) representação da informação situacional. O processo como um todo visa à construção e melhoria contínua das informações, para que possam atender ao domínio situado no contexto de gerenciamento de emergências, como o controle e combate a incêndios.

A definição dos critérios de qualidade deve ser dada por tomadores de decisão especialistas no domínio de gerenciamento de emergências, no caso, os bombeiros. Essa definição pode ser realizada mediante uma análise de tarefas dirigida por objetivos (*Goal-Driven Task Analysis* - GDTA), técnica proposta por Endsley (2016) e aplicada em Oliveira *et al*. (2017) e Silva *et al*. (2018), que tem por fim revelar as tarefas a serem realizadas, as decisões a serem tomadas e as informações necessárias à tomada de decisões.

A técnica propõe a aplicação de um questionário cujas questões possam ser classificadas segundo a percepção e compreensão que se obtém diante de uma situação de incêndio, por exemplo. Deve-se questionar sobre quais são as tarefas diárias realizadas, quais as decisões tomadas para realizar tais tarefas e, por fim, quais informações são necessárias para tomar cada decisão.

De acordo com os resultados obtidos na coleta, um especialista do domínio de gerenciamento de emergências pode definir quais dimensões de qualidade serão relevantes, tais como: atualidade, consistência, relevância e completude. À vista disso, torna-se possível iniciar o processo de adaptação da ontologia de domínio desenvolvida por Silva *et al*. (2018), conforme a figura 1, por meio do reuso de outras ontologias, que devem possuir a característica temporal.

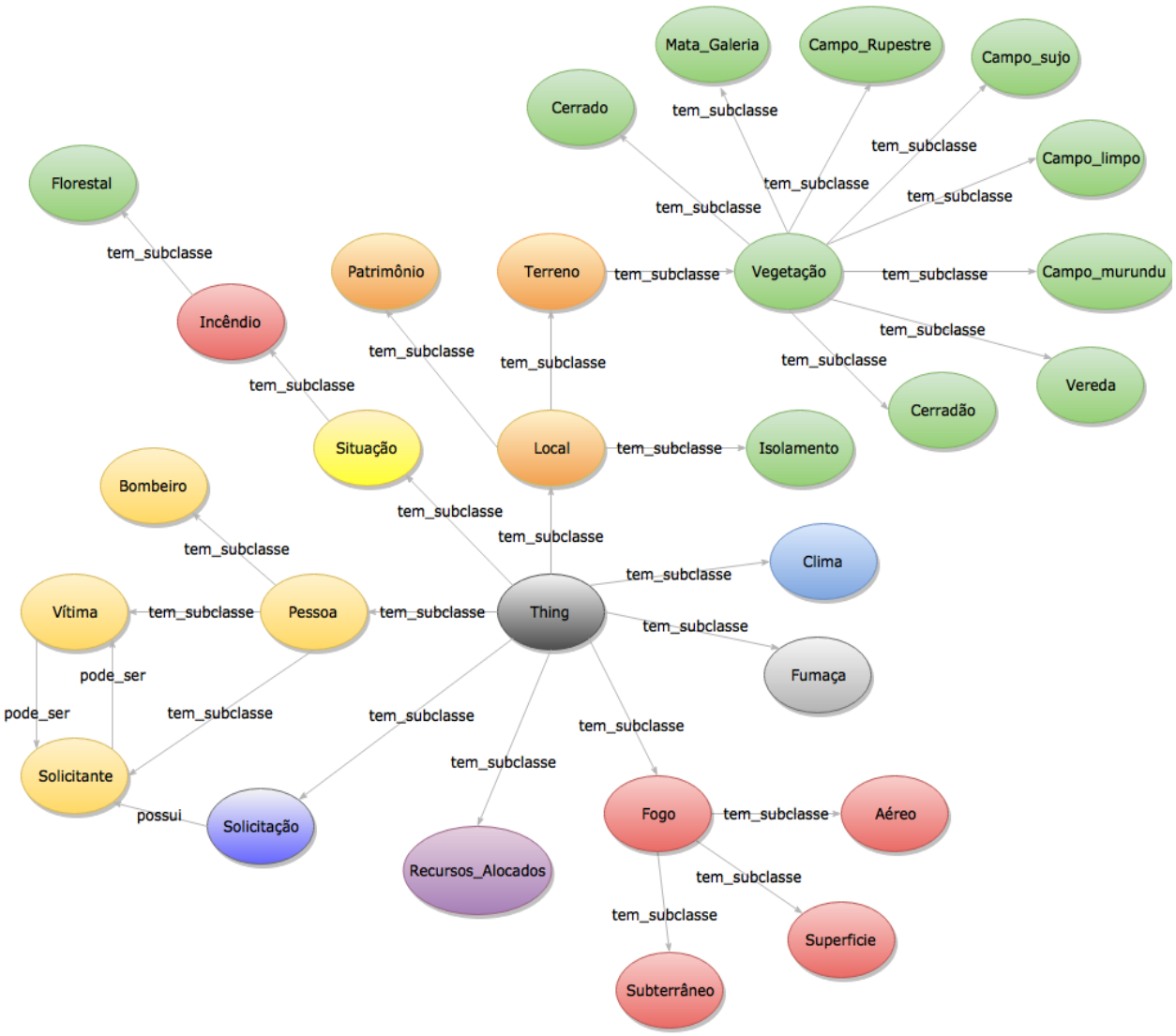
A ontologia Time Ontology in OWL[1], mencionada por Tao *et al*. (2010), é útil diante da sua aplicação na ontologia desenvolvida por Silva *et al*. (2018), pois ela não possui a dimensão temporal em seu escopo. Os relacionamentos temporais presentes na ontologia possibilitam a criação de uma linha do tempo capaz de fornecer novos insumos informacionais por meio de inferências.

Ademais, pode-se incluir a SWRL Time Ontology[2], desenvolvida pela Universidade Stanford, uma vez que ela conta com um vocabulário derivado da álgebra de intervalo de Allen, isto é, termos vinculados às relações temporais, como "igual", "antes", "depois", "toca", "sobrepõe", "durante", "inicia" e "finaliza". Essa ontologia permite a interoperabilidade concernente aos operadores temporais de Allen utilizados junto aos relacionamentos temporais definidos na ontologia proposta pela W3C. Os operadores temporais utilizados pela SWRL Temporal Ontology permitem que os relacionamentos temporais das classes da Time Ontology in OWL existam. Ou seja, é possível relacionar duas instâncias da classe Evento por intermédio do operador "depois", resultando na tripla em que a primeira instância ocorreu após a segunda instância.

[1] https://www.w3.org/TR/owl-time/

[2] https://github.com/protegeproject/swrlapi/wiki/SWRLTemporalOntology

Figura 2 – Ontologia de domínio



Fonte: Silva *et al.* (2018).

A ontologia da W3C, a Time Ontology in OWL, conta com duas principais classes: Evento e Tempo. A primeira trata de qualquer tipo de ocorrência, estado, percepção, procedimento, sintoma ou situação que ocorra em uma linha do tempo. A segunda é dividida em outras quatro classes: Instante, Intervalo, Fase e Período.

A classe Instante refere-se a um ponto específico de tempo dentro de uma linha temporal, na qual existem fatores granulares, como data (ano, mês e dia) e horário (hora, minuto e, se necessário, segundo). Tais granularidades permitem que a linha do tempo seja representada e recuperada de maneira correta pela ontologia, além de auxiliar nos processos de inferências, para que novos conhecimentos temporais sejam descobertos.

A classe Intervalo representa a duração de tempo, ou seja, há um relacionamento de início e fim. Cada parte do relacionamento torna-se uma instância de Instante.

A classe Fase representa cada ocorrência de um intervalo repetido, também tendo início e fim. Para finalizar, a classe Período específica a medida de frequência que uma Fase repete.

Toda informação que remete à horário, independente de qual classe temporal for, deve ser representada pela classe Duração, que deve conter a unidade de tempo utilizada junto ao seu respectivo valor. A unidade de tempo é dada pelo fator granular mencionado acima, isto é, pode ser "ano", enquanto seu valor é "2019", por exemplo.

Apesar dessas classes estarem presentes no modelo semântico, elas não cumprirão seus objetivos se não houver um relacionamento consistente entre elas. Portanto, o relacionamento temporal se dá entre duas instâncias de Evento ou de Evento com alguma instância de Tempo, como evidenciam os operadores temporais utilizados pela SWRL Temporal Ontology, à medida que os relacionamentos temporais das classes da Time Ontology in OWL possibilitam relacionar duas instâncias da classe Evento (e1 e e2, respectivamente) por meio do operador "depois", resultando na tripla "e1 ocorreu depois de e2", por exemplo.

Para validar o modelo proposto, bem como as adaptações na ontologia e a inferência temporal, é necessária uma prova de conceito, que será apresentada na próxima seção.

PROVA DE CONCEITO

A ontologia proposta por Silva *et al*. (2018) conta com as classes Situação, Solicitação, Local, Clima, Pessoa, entre outras. O principal destaque é para a classe de Solicitação, que remete a um evento de uma situação de emergência. Essa classe pode ser considerada equivalente à classe de Evento da Time Ontology in OWL, uma vez que uma solicitação ocorre em uma linha do tempo.

A implementação da dimensão temporal segue as recomendações dadas por Tao *et al*. (2010, 2011), baseando-se no estudo de caso de Silva *et al*. (2018). A nível de instâncias, duas foram criadas a partir da classe Solicitação. A primeira solicitação foi emitida por um cidadão, é do tipo alerta, possui confiabilidade do emissor e o horário da denúncia foi às 14h23. A segunda foi emitida por um bombeiro, também é do tipo alerta e possui confiabilidade do emissor, sendo o horário da denúncia às 14h25.

Ambas as solicitações podem corresponder à classe de Evento da Time Ontology in OWL, portanto, o relacionamento temporal pode ser criado ao vincular uma instância de Instante, que concerne a um momento de tempo.

A unidade de tempo das duas instâncias deve ser especificada como horário e os atributos de tempo precisam ter seus valores normalizados para serem interpretados por máquinas, desse modo, os horários "14h23" e "14h25" devem ser normalizados para "14:23" e "14:25", respectivamente, conforme a ISO 8601 - Formato de Data e Hora.

A representação das instâncias das situações e do relacionamento temporal é apresentada pela figura 3A e figura 3B.

Na figura 3A os objetos com preenchimento na cor laranja representam as classes da ontologia de domínio proposta por Silva *et al*. (2018), enquanto os objetos com preenchimento na cor roxa representam as instâncias dessas classes, que, de sua parte, remetem a esta prova de conceito. As setas contínuas rosas, referindo-se a qual classe pertence uma instância específica, enquanto as linhas tracejadas laranjas referem-se ao relacionamento entre instâncias de outras classes.

Por sua vez, na figura 3B os objetos com preenchimento na cor verde são os valores das propriedades relacionadas às instâncias. A linha contínua azul diz respeito à subclasse e a contínua verde remete ao vínculo entre propriedade e valor específico. Para fins de destaque, a inclusão da dimensão temporal é representada pelo todo que se encontra dentro da borda vermelha.

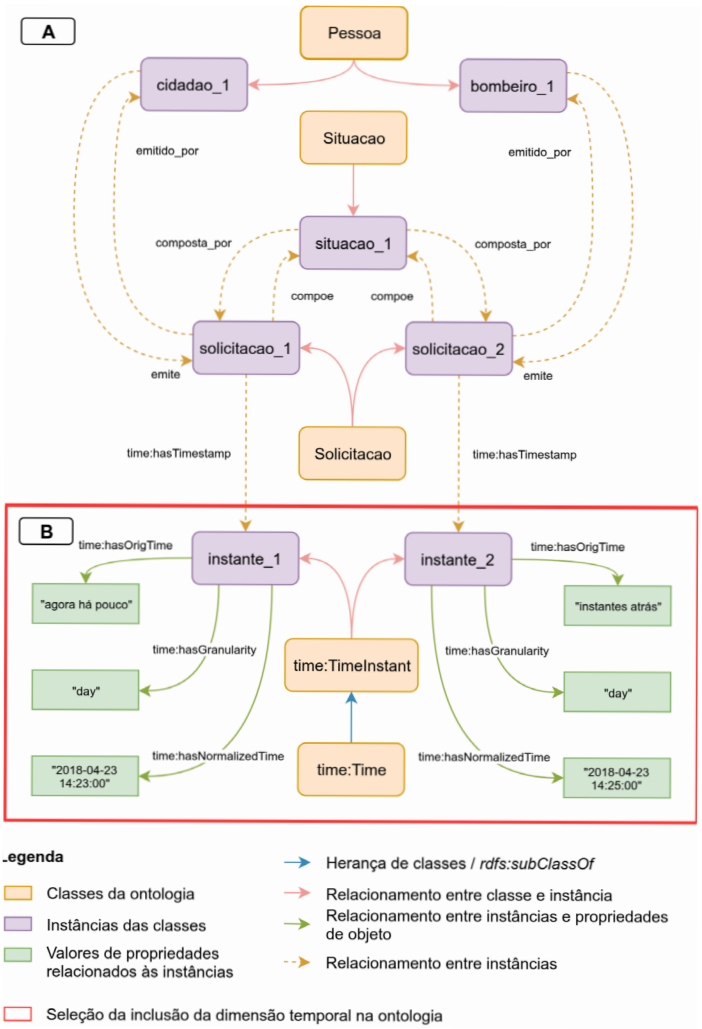
De acordo com o estudo de Silva *et al.* (2018), ocorreram duas solicitações ("solicitacao_1" e "solicitacao_2"), que concernem à mesma situação ("situacao_1"). Entretanto, caso não houvesse a dimensão, não seria possível dizer ontologicamente qual evento ocorreu antes do outro, por exemplo.

Dada a inclusão da dimensão temporal, há a possibilidade de inferir novos conhecimentos temporais a partir dos novos relacionamentos criados, como, por exemplo, a Solicitação 1 ter ocorrido antes da Solicitação 2, utilizando a propriedade "time:hasNormalizedTime" das instâncias "instante_1" e "instante_2", sendo expressada pela regra "swrl:before(?instante_1, ?instante_2)".

As métricas qualitativas e quantitativas podem ser afetadas, uma vez que uma das dimensões de qualidade passou a ter relacionamentos ternários passíveis de novas mensurações.

A partir dessa adequação e do reuso das ontologias mencionadas acima, a ontologia proposta por Silva *et al.* (2018) torna-se capaz para suportar os demais relacionamentos temporais.

Figura 3 – (A) Representação da ontologia sem a dimensão temporal



Fonte: Silva *et al.* (2018).

CONSIDERAÇÕES FINAIS

A ciência da informação, enquanto disciplina que investiga as propriedades e o comportamento da informação, preocupa-se com a resolução de problemas da efetiva comunicação, interação e integração do conhecimento e seus respectivos elementos, de acordo com as necessidades informacionais. A informação, como objeto de estudo da Ciência da Informação, pode proporcionar um ambiente interdisciplinar, estimulando discussões sob diferentes perspectivas e áreas envolvidas.

Nesse prisma, a recuperação da informação está intimamente relacionada à perspectiva apresentada acima. Isso se deve a um dos objetivos da Recuperação da Informação, que é proporcionar melhores insumos informacionais àqueles que precisem recuperar quaisquer informações, a fim de extinguir suas necessidades informacionais.

Suprir as necessidades informacionais de indivíduos situados no domínio de gerenciamento de emergências é um processo que se torna possível a partir do uso de modelos semânticos, tais como as ontologias, que possuem alta capacidade de inferir novos descobrimentos informacionais. É cabível ressaltar que, devido às características ímpares desse ambiente informacional, é demandado que as informações expressem índices qualitativos fidedignos.

Diante disso, este trabalho teve, como objetivo, as melhorias dos processos informacionais da recuperação da informação com base em um modelo de análise temporal em contexto semântico de gerenciamento de emergências. Esse modelo conta com uma definição do processo informacional que permita a interação e integração da recuperação da informação, a qualidade da informação, a semântica e o espectro temporal.

Ademais, a temporalidade possibilita a análise informacional pela perspectiva de uma linha do tempo, apresentando a evolução das informações no decorrer do tempo.

Esse aspecto, atrelado à qualidade da informação, possibilita que as informações recuperadas por indivíduos sejam mais ricas, sirvam de insumos para a tomada de decisão e, consequentemente, para suprir suas necessidades informacionais.

As regras presentes no domínio podem ser consideradas fatores limitantes, entretanto elas são as responsáveis por guiar a estrutura do processo informacional, a exemplo da criação de uma regra inferencial a ser aplicada na ontologia. Tais regras devem ser elaboradas e validadas por especialistas antes de serem implementadas, para evitar erros e incertezas que possam prejudicar todo o aporte informacional.

REFERÊNCIAS

BATINI, C. *et al*. Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, [*s. l.*], v. 41, n. 3, artigo 16, 52 p., 2009. DOI: https://doi.org/10.1145/1541880.1541883.

BOTEGA, L. C. *Modelo de fusão dirigido por humanos e ciente de qualidade de informação.* 2016. 247 f. Tese (Doutorado em Ciência da Computação) - Universidade Federal de São Carlos, São Carlos, 2016.

BOTEGA, L. C. *et al*. Methodology for data and information quality assessment in the context of emergency situational awareness. *Universal Access in the Information Society*, [*s. l.*], v. 16, n. 4, p. 889-902, 2017. DOI: https://doi.org/10.1007/s10209-016-0473-0.

BOTEGA, L. C. *et al*. Quantify: an information fusion model based on syntactic and semantic analysis and quality assessments to enhance situation awareness. In: BOSSÉ, E.; ROGOVA, G. L. (org.). *Information Quality in Information Fusion and Decision Making*, [*s. l.*]: Springer Cham, 2019, p. 563-586. ISBN 978-3-030-03643-0.

BUCKLAND, M. K. Information as thing. *Journal of the American Society for information science (JASIS)*, [*s. l.*], v. 45, n. 5, p. 351-360, 1991.

CALAZANS, A. T. S. Qualidade da informação: conceitos e aplicações. *TransInformação,* [*s. l.*], v. 20, n. 1, p. 29-45, 2008. DOI: https://doi.org/10.1590/S0103-37862008000100003.

ENDSLEY, M. R. *Designing for situation awareness: an approach to user-centered design*. 2. ed. [*s. l.*]: CRC press, 2016. 396 p. ISBN 978-1420063554.

GÓMEZ-PÉREZ, A.; BENJAMINS; R. Overview of knowledge sharing and reuse components: Ontologies and problem-solving methods. In: IJCAI-99 WORKSHOP ON ONTOLOGIES AND PROBLEM-SOLVING METHODS (KRR5), 1999, Stockholm, Sweden: Faculty of Social and Behavioural Sciences (FMG), 1999. Disponível em: https://dare.uva.nl/search?identifier=b6e475ac-2649-4f68-a181-4e42f9eb6ce7. Acesso em: 11 mar. 2021.

HASANUZZAMAN, M. *et al*. Propagation strategies for building temporal ontologies. In: CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 14, 2014, Gothenburg, Sweden: Association for Computational Linguistics. *Short Papers*, v. 2, p. 6-11, 2014. DOI: 10.3115/v1/E14-4002.

KOKAR, M. M.; ENDSLEY, M. R. Situation awareness and cognitive modeling. *IEEE Intelligent Systems,* [*s. l.*], v. 27, n. 3, p. 91-96, 2012. DOI: 10.1109/MIS.2012.61.

LIU, L.; CHI, L. Evolutional data quality: a theory-specific view. In: INTERNATIONAL CONFERENCE ON INFORMATION QUALITY (ICIQ-02), 7, 2002, [*s. l.*]: ICIQ, *Research Paper*, 2002, p. 292-304. Disponível em: http://mitiq.mit.edu/ICIQ/Documents/IQ%20Conference%202002/Papers/EvolutionalDataQualityATheorySpecificView.pdf. Acesso em: 11 mar. 2021.

MATHEUS, C. J.; KOKAR, M. M.; BACLAWSKI, K. A core ontology for situation awareness. In: INTERNATIONAL CONFERENCE ON INFORMATION FUSION, 6, 2003, Cairns, QLD, Australia, *Anais* […], Cairns, QLD, Australia: IEEE, 2003, p. 545-552. DOI: 10.1109/ICIF.2003.177494.

MELO, J. O. de S. F.; BOTEGA, L. C.; SANTARÉM SEGUNDO, J. E. Metodologia de avaliação de qualidade para dados conectados. *Informação & Tecnologia (ITEC)*, Marília/João Pessoa, v. 4, n. 2, p. 80-101, jul./dez 2017. DOI: https://doi.org/10.22478/ufpb.2358-3908.2017v4n2.40539.

NEHMY, R. M. Q.; PAIM, I. A desconstrução do conceito de "qualidade da informação". *Ciência da Informação*, Brasília, v. 27, n. 1, p. 36-45, jan./abr 1998. DOI: https://doi.org/10.1590/S0100-19651998000100005.

OKEYO, G.; CHEN, L.; WANG, H. Combining ontological and temporal formalisms for composite activity modelling and recognition in smart homes. *Future Generation Computer Systems*, [*s. l.*], v. 39, p. 29-43, 2014. DOI: https://doi.org/10.1016/j.future.2014.02.014.

OLETO, R. R. Percepção da qualidade da informação. *Ciência da informação*, Brasília, v. 35, n. 1, p. 57-62, jan./abr. 2006. DOI: http://dx.doi.org/10.1590/S0100-19652006000100007.

OLIVEIRA, A. C. M. *et al*. Crowdsourcing, data and information fusion and situation awareness for emergency management of forest fires: the project DF100Fogo (FDWithoutFire). *Computers, Environment and Urban Systems*, [*s. l.*], v. 77, 2017. DOI: https://doi.org/10.1016/j.compenvurbsys.2017.08.006.

OLSON, J. E. *Data quality: the accuracy dimension*. San Francisco: Morgan Kaufmann, 2003. 312 p.

PAZ-TRILLO, C.; WASSERMANN, R.; BRAGA, P. P. An information retrieval application using ontologies. *Journal of the Brazilian Computer Society*, Campinas, v. 11, n. 2, p. 17-31, 2005. DOI: http://dx.doi.org/10.1007/BF03192373.

PEREIRA JUNIOR, V. A.; PEREIRA, G. M. C.; BOTEGA, L. C. Towards a Process for Criminal Semantic Information Fusion to Obtain Situational Projections. In: HAYNES, D.; VERNAU, J. (Ed.). *The Human Position in an Artificial World: creativity, ethics and ai in knowledge organization.* Londres: Ergon-Verlag, 2019. p. 51-72. DOI: https://doi.org/10.5771/9783956505508-51.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. *Perspectivas em ciência da informação,* Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/235. Acesso em: 11 mar. 2021.

SILVA, J. N. *et al*. Desenvolvimento de ontologia ciente de qualidade de informações para o domínio de gerenciamento de emergências. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, Florianópolis, v. 23, n. 53, p. 184-200, 2018. DOI: 10.5007/1518-2924.2018v23n53p182.

TAO, C. *et al*. CNTRO: a semantic web ontology for temporal relation inferencing in clinical narratives. In: AMIA ANNUAL SYMPOSIUM PROCEEDINGS. *American Medical Informatics Association.* p. 787-791, 2010.

TAO, C.; SOLBRIG, H. R.; CHUTE, C. G. CNTRO 2.0: a harmonized semantic web ontology for temporal relation inferencing in clinical narratives. *AMIA summits on translational science proceedings,* v. 2011, p. 64-68, 2011.

WIEGAND, N.; GARCÍA, C. A task-based ontology approach to automate geospatial data retrieval. *Transactions in GIS*, v. 11, n. 3, p. 355-376, 2007. DOI: 10.1111/j.1467-9671.2007.01050.x

Interlocuções bibliográficas e epistemológicas entre a ciência de dados e a ciência da informação

Jorge Henrique Cabral Fernandes
Doutor, Universidade Federal de Pernambuco (UFPE), Recife, PE, Brasil
Professor Adjunto IV, Universidade de Brasília (UnB), Brasília, DF, Brasil
http://lattes.cnpq.br/7151669913805328
E-mail: jhcf@unb.br

Submetido em: 30/04/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021.

RESUMO

Diante da grande evidência alcançada recentemente por métodos, ferramentas e práticas da ciência de dados, este artigo verifica a possibilidade de que existam novos fundamentos em uma suposta ciência chamada de Ciência de Dados e de que forma ela impactaria ou seria impactada pela Ciência da Informação. A metodologia é exploratória e usa pesquisa bibliográfica baseada em análise de redes de co-ocorrências e de acoplamento bibliográfico envolvendo os termos "data science", "information science" e "library and information science". As interlocuções e contribuições dos dois campos de práticas, evidenciadas pelas análises produzidas, bem como de algumas implicações epistemológicas, apontam não apenas para uma oportunidade de intercâmbios, mas para uma necessidade de que eles ocorram com a maior brevidade.

Palavras-chave: Big data. Biblioteconomia. Análise de co-ocorrência. Acoplamento bibliográfico. Desenvolvimento profissional.

Bibliographic and epistemological exchanges between data science and information science

ABSTRACT

In view of the great recent evidence achieved by methods, tools and practices of data science, the article verifies the possibility that there are new foundations in a supposed science called Data Science, and how it would impact or be impacted by Information Science. The methodology is exploratory, and uses bibliographic research based on analysis of networks of co-occurrences and bibliographic coupling involving the terms "data science", "information science" and "library and information science". The exchanges and contributions of the two fields of practices, evidenced by the analyzes produced, as well as some epistemological implications, point not only to an opportunity for exchanges, but to a need for them to occur as soon as possible.

***Keywords**: Big data. Libray and information science. Analysis of co-occurrences. Bibliographic coupling. Professional development.*

Interlocuciones bibliográficas y epistemológicas entre ciencia de datos

y ciencia de la información

RESUMEN

En vista de la gran evidencia alcanzada recientemente por los métodos, herramientas y prácticas de la ciencia de datos, el artículo verifica la posibilidad de que haya nuevas bases en una supuesta ciencia llamada Ciencia de Datos, y de qué manera impactaría o sería impactada por la Ciencia de Información. La metodología es exploratoria y utiliza investigación bibliográfica basada en análisis de redes de coincidencias y acoplamiento bibliográfico que involucra los términos "ciencia de datos", "ciencia de la información" y "biblioteca y ciencia de la información". Las interlocuciones y contribuciones de los dos campos de prácticas, evidenciadas por los análisis producidos, así como algunas implicaciones epistemológicas, apuntan no solo a una oportunidad para intercambios, sino a la necesidad de que ocurran lo antes posible.

Palabras clave*: Big data. Biblioteconomía. Análisis de concurrencia. Acoplamiento bibliográfico. Desarrollo profesional.*

INTRODUÇÃO

É inegável a intensa valorização recente da área de atividade profissional associada ao termo "Ciência de Dados" (DAVENPORT; PATIL, 2012). Seria a ciência de dados uma área científica em si ou trata-se apenas de uma atividade passageira que, devido à pouca consistência, coesão ou abrangência de propósito, estaria fadada a desaparecer, sendo absorvida por áreas já estabelecidas, como computação, estatística, ciência da informação ou administração? Traria, a área de ciência de dados, novos fenômenos dignos de investigação nos campos filosóficos, epistemológicos e científicos da atividade humana, bem como numa prática profissional distintiva?

Embora exista uma razoável confusão sobre o que de fato faz um cientista de dados e sobre o que seja a ciência de dados, pode-se recorrer à pesquisa em bases de dados bibliográficas para que se tenha uma compreensão da atividade. Também é admissível o aprofundamento filosófico-epistemológico da ciência de dados, a fim de identificar se, de fato, um conjunto distintivo de fenômenos de relevância científica pode ser agregado a essa suposta nova ciência. Diante disso, este artigo busca verificar se existem novos fundamentos a explorar na presumida ciência de dados e, se for esse o caso, quais seriam as interlocuções possíveis dessa nova ciência com a já estabelecida ciência da informação.

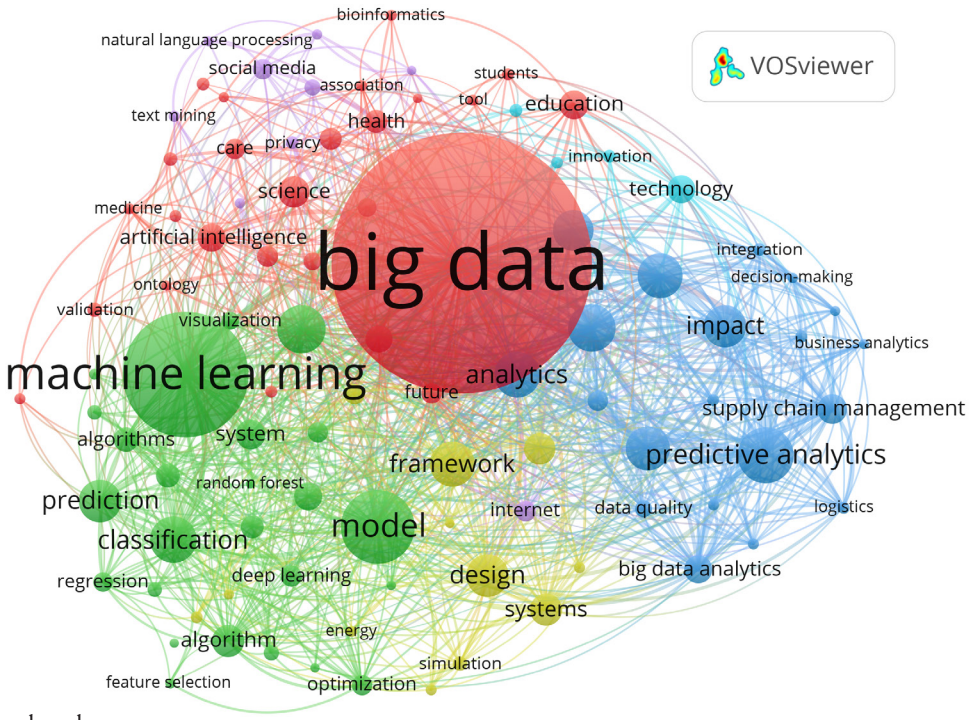
METODOLOGIA

Este trabalho adota, como metodologia, a análise de registros bibliográficos obtidos na *Core Collection* da base *Web of Science* (WoS), ofertada por meio do portal de periódicos da CAPES e associada às publicações de artigos científicos até abril de 2020. Também emprega a ferramenta VOSViewer (VAN ECK; WALTMAN, 2010) para geração de redes e análises de co-ocorrência de palavras-chave e acoplamento bibliográfico.

Nas redes de co-ocorrência de palavras-chave, duas palavras-chave são vinculadas uma à outra, caso apareçam simultaneamente em um documento. Quanto maior a quantidade de documentos em que esse par ocorre, maior é a associação entre essas duas palavras-chave. A suposição é de que quanto maior for a co-ocorrência dessas palavras, mais afinidades elas possuem entre si.

Nas redes de acoplamento bibliográfico, considera-se que duas unidades de análise (documentos, fontes de publicação, autores, organizações ou países) estão vinculadas a partir do momento em que citam a(s) mesma(s) referência(s) bibliográfica(s). Nessa perspectiva, quanto maior a quantidade de referências comuns citadas, maior é a suposição de que as unidades de análise tratam do mesmo assunto.

Figura 1 – Rede de co-ocorrência de palavras-chave em publicações em ciência de dados



Fonte: Elaborada pelo autor.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

A CONDIÇÃO BIBLIOGRÁFICA ATUAL DA *DATA SCIENCE* (CIÊNCIA DE DADOS)

A pesquisa na base WoS, conforme as abordagens apresentadas na metodologia, obteve, entre 2011 e 2020, 1.880 registros de artigos científicos associados aos termos "data science" ou "data-science". A rede de co-ocorrência das 96 palavras-chave mais co-citadas nesses 1.880 registros, gerada pelo software VOSViewer, é apresentada na figura 1.

Para a apresentação de uma rede com, no máximo, 100 elementos, aumentou-se gradualmente o limite mínimo de co-citações para cada palavra-chave, alcançando-se a quantidade de 98 palavras-chave com um mínimo de 16 co-ocorrências. Foram excluídos os termos "data science" e "data-science" e tamanho das palavras-chave na figura 1 é proporcional à quantidade de vezes em que aparecem. Por fim, a rede da figura 1 é composta de seis *clusters* de palavras, delimitando os seguintes temas característicos da ciência de dados, no geral:

Big data: termo mais proeminentemente associado à ciência de dados e ao maior *cluster* da rede, composto de 26 termos. O *big data* representa a oportunidade que é perdida ou aproveitada pelas organizações humanas, na medida em que geram ou têm acesso a dados em grande volume, grande variedade, produzidos em alta velocidade, com ampla variação em veracidade e elevado potencial de produção de valor (WANG, 2018). O *cluster* dominado pelo *big data* está fortemente relacionado a todos os outros. Nesse *cluster*, ainda aparecem termos como "education", "artificial intelligence", "statistics", "health" e "data analysis", sugerindo a vinculação do *big data* à aplicação da inteligência artificial e de técnicas estatísticas para análises e sínteses de modelos nos campos de aplicação da educação e saúde;

Machine learning: segundo termo mais evidente em associação com a ciência de dados, domina o segundo maior *cluster* da rede, com 25 termos. A aprendizagem de máquina está associada aos termos "model", "prediction", "classification", "algorithms" e compreende basicamente o uso de dados do passado para treinar sistemas algorítmicos e computacionais para identificação de padrões em dados, tendo por fim criar modelos de predição de eventos futuros ou de classificação de situações presentes, baseados em dados recém-disponíveis. Sheble (2016) ilustra situações em que a aprendizagem de máquina pode facilitar a seleção de artigos mais relevantes em pesquisas bibliográficas. Kazakci (2015) ilustra o modo como a aprendizagem de máquina pode auxiliar na identificação de sinais relevantes em um estudo para detecção de partículas de alta energia em física;

Predictive analytics: termo mais evidente do terceiro mais proeminente *cluster* do mapa de co-ocorrência de palavras-chave associadas às publicações em ciência de dados. A "analytics" preditiva consiste na aplicação da ciência de dados pela gestão ("management"), inclusive na cadeia de suprimento de empresas ("supply chain management"), visando ao melhor desempenho ("performance") e gerando grande impacto ("impact") nos negócios por meio da informação ("information") para desenho de estratégia e da obtenção de vantagens competitivas em sistemas logísticos. Ortiz-Repiso, Greenberg e Calzada-Prado (2018) vinculam o uso de *analytics* em empresas à imensa valorização da atividade profissional dos chamados cientistas de dados;

Framework: termo mais evidente associado ao quarto *cluster* de palavras-chave em ciência de dados. Sua vinculação aos termos "systems", "design", "data analytics", "simulation", "dynamics" e "uncertainty", evidencia a necessidade de desenvolvimento de arcabouços computacionais (frameworks) de sistemas (de informação) baseados em técnicas de simulações de sistemas dinâmicos, com vistas a reduzir a incerteza na tomada de decisão ("decision making").Kazakci (2015) apresenta de que forma o *design* coletivo de uma aplicação de aprendizagem de máquina é capaz de produzir inovações de forma muito ágil;

Social media: termo mais evidente do segundo menor *cluster*, com nove itens, e que destaca o intenso uso de abordagens em ciência de dados para analisar dados gerados pelas mídias sociais, a exemplo do *Twitter*, por meio de processamento de linguagem natural e mineração de textos. Essa utilização apresenta implicações éticas, ligadas a ameaças à privacidade na sociedade. A reduzida dimensão desse *cluster* evidencia a necessidade de avanços na compreensão dos problemas relativos à exploração de dados da sociedade sem o devido cuidado ético; e

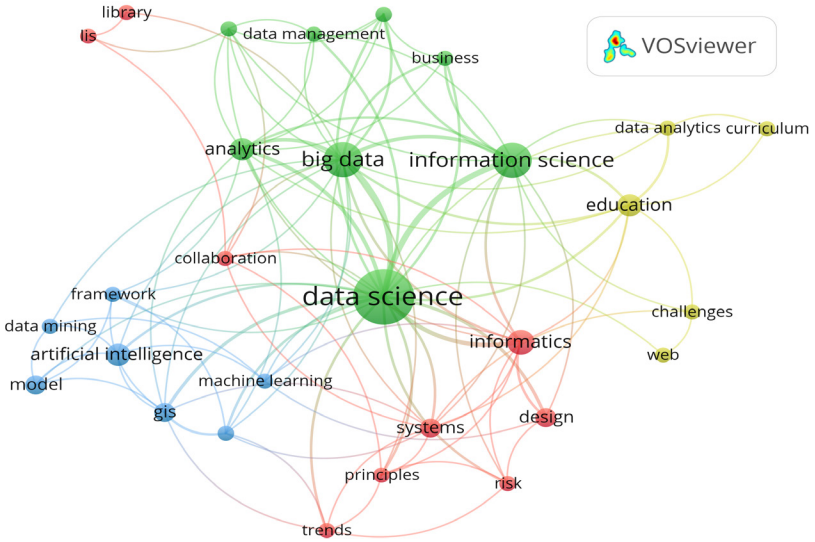
Tecnologia: termo que representa o menor dos *clusters* de palavras-chave e está associada à produção de inovações. A ciência de dados, representada pela aprendizagem de máquina e pela inteligência artificial, é uma grande indutora de inovações.

A partir da análise dos *clusters* pode-se sintetizar a atividade de ciência de dados como sendo, atualmente, aquela que busca explorar as oportunidades do *big data* por meio da criação de sistemas computacionais de análise, inclusive de natureza preditiva, visando ao desenvolvimento de inovações tecnológicas para empresas em um ambiente competitivo, especialmente de cadeia de suprimentos, ou para usos em áreas como saúde e educação, mediante aplicação de técnicas como aprendizado de máquina e inteligência artificial. A atividade dos cientistas de dados, quando aplicada às mídias sociais, pode gerar implicações éticas de ameaça à privacidade.

INTERLOCUÇÃO BIBLIOGRÁFICA DA CIÊNCIA DA INFORMAÇÃO COM A CIÊNCIA DE DADOS

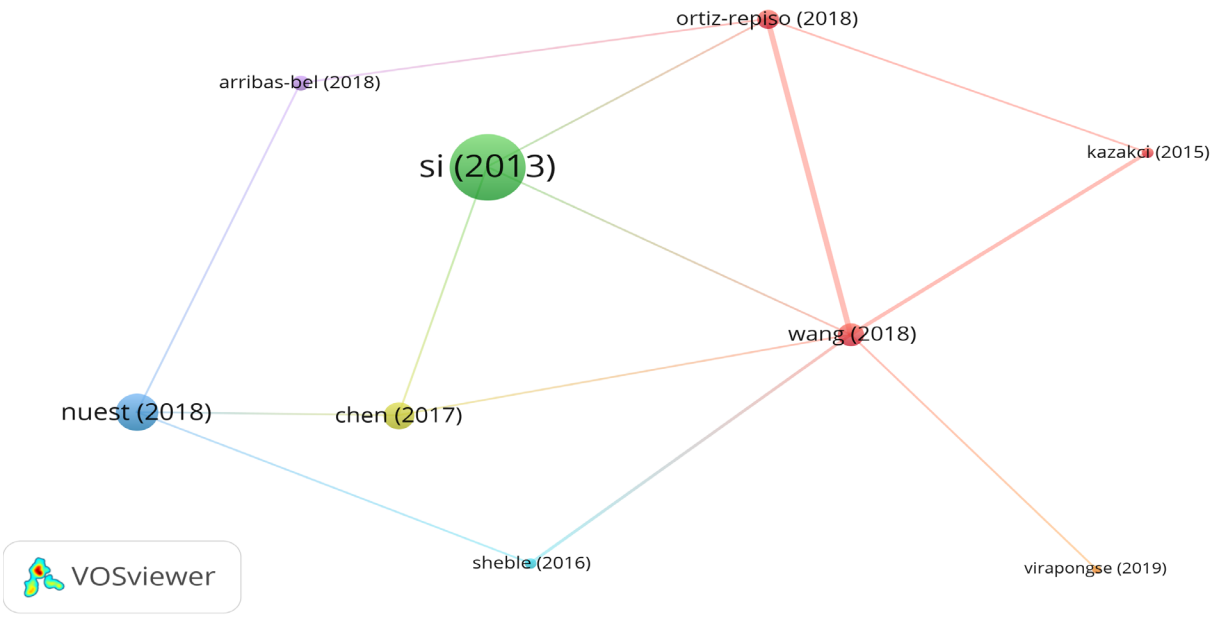
Uma segunda pesquisa na base de dados bibliográfica foi efetuada pelo autor, dessa vez usando a *string* de busca "(('data science' or data-science) and ('information science' or 'library science' or lis))", aplicável todos os anos. Foi retornado um pequeno conjunto de documentos em todos os tipos disponíveis na base, 34 ao todo, sendo 25 artigos de *journal*, seis *proceeding papers* e três revisões.

Figura 2 – Mapa de co-ocorrência de palavras-chave que combinam ciência de dados e ciência da informação ou "library science"



Fonte: Elaborada pelo autor.

Figura 3 – Mapa de acoplamento bibliográfico entre itens que combinam ciência de dados e ciência da informação ou "library science"



Fonte: Elaborada pelo autor.

ANÁLISE DE CO-OCORRÊNCIAS

A figura 2 apresenta a rede de co-ocorrência de palavras-chave usadas nos 34 documentos que associam ciência de dados e ciência da informação, e que possuem pelo menos duas associações. A rede possui quatro *clusters*, sendo que os termos "data science", "information science" e "big data" continuam a dominar o *cluster* principal, que possui oito elementos. Além disso, esse *cluster* apresenta ainda a ocorrência de termos ligados à gestão de dados e a negócios.

O *cluster* mais numeroso da rede, com nove elementos, apresentado em vermelho, possui o termo "informatics" como mais dominante, apesar de próximo dos termos "design" e "systems". Desse mesmo *cluster*, fazem parte os termos ligados às "library sciences" e "lis", que, embora agrupados com "informatics", se encontram posicionados próximos ao "big data", "data management". Nesse *cluster*, a ocorrência de vários termos sugere preocupações com fenômenos em contextos sociais, tais como riscos ("risks"), colaborações ("collaboration"), princípios ("principles") e tendências ("trends").

O terceiro mais importante *cluster* da rede apresenta predominância de termos de origem computacional e estatística, como inteligência artificial, aprendizagem de máquina, mineração de dados e sistemas de informação georreferenciados. Termos ligados a fatores humanos ou sociais estão ausentes nesse *cluster*.

O último e menor dos *clusters* de palavras-chave que aparecem na relação entre ciência de dados e ciência da informação enfatiza "education" e seus desafios ("challenges"), e o "curriculum" de "data analytics" apresenta-se como foco. Os termos sugerem a existência de desafios à incorporação de temas como *data science*, *big data* e *analytics* à ciência da informação ou, de outro modo, a possibilidade de também existirem desafios educacionais relacionados à incorporação de temas da ciência da informação à ciência de dados.

ACOPLAMENTO BIBLIOGRÁFICO

A fim de melhor detalhar as principais discussões dos pontos de interseção entre a ciência de dados e a ciência da informação, pode-se analisar a rede de acoplamento bibliográfico dos 34 documentos anteriormente recuperados, buscando-se observar de que forma se associam os itens mais relevantes na interlocução entre a ciência de dados e a ciência da informação. Dentre esses 34 documentos, 17 são citados pelos demais e apenas nove formam um componente de rede, isto é, fazem citação de bibliografia comum. A rede é apresentada na figura 3.

A leitura e análise dos nove documentos acoplados pela bibliografia é apresentada a seguir, em busca de aprofundar e sugerir os futuros caminhos de desenvolvimento de uma interlocução entre a ciência de dados e a ciência da informação. Os artigos serão apresentados na ordem de maior centralidade na rede apresentada na figura 3.

Twinning data science with information science in schools of library and information science, de Wang (2018), apresenta a ciência de dados como uma nova corrente vital na educação em "library and information science", onde a ciência de dados seria uma "irmã gêmea" da ciência da informação, com um grande número questões em comum. Wang descreve de que forma os conceitos da ciência da informação contribuiriam para aprimorar a ciência de dados. Vinculado ao Departamento de Administração da Universidade Normal de Tianjin, na China, mas citando avanços da integração disciplinar nos EUA, Wang (2018) argumenta que as escolas de informação devem se tornar ambidestras na integração de ambas as "ciências". Nesse sentido, trariam contribuições significativas à ciência de dados, os seguintes aspectos oriundos da ciência da informação (WANG, 2018):

a) Concepção de dados nos sentidos históricos e pragmáticos, em complemento à visão positivista dominante em ciência de dados, que, do ponto de vista analítico, tende a considerar dados como sendo objetivos e neutros;

b) Controle da qualidade de dados, visando promover técnicas para aumentar a precisão, acurácia e credibilidade de dados, especialmente quando tratando-se de *big data*;

c) Organização bibliotecária de dados, com vistas a promover, em ciência de dados, o uso de técnicas e ferramentas voltadas a coleção, preservação, curadoria, governança, metadados, compartilhamento, visualização, suporte à análise, avaliação de qualidade, referenciação, citação e literacia de dados; e

d) Estudo das teorias da documentação, com o objetivo de trazer para a ciência de dados a percepção de que ela trabalha com documentos e que documentos apresentam três dimensões essenciais (BUCKLAND, 2016): dimensão tecnológica de seu suporte físico; dimensão cognitiva da relação mental dos indivíduos com os documentos; e dimensão social, relacionada aos papéis econômicos e políticos dos documentos.

Enquanto contribuição inversa da ciência dos dados para a ciência da informação, o autor aponta para o *big data* como gerador de demandas por mudanças na ciência da informação, mas sem maiores detalhamentos.

A cross-institutional analysis of data-related curricula in information science programmes: A focused look at the iSchools, de Ortiz-Repiso, Greenberg e Calzada-Prado (2018), apresenta uma análise da ênfase em estudos sobre dados em 597 cursos que conferiram graus em 65 escolas do consórcio internacional das *i-Schools* (*information schools*), no período de fevereiro de 2016. O levantamento indicou que 14% dos programas analisados enfatizavam estudos sobre dados, sendo que 11% desses estudos voltavam-se ao *data science* em geral ou, mais especificamente, ao *big data analytics*; e 3% eram dirigidos à curadoria de dados digitais.

The cultivation of scientific data specialists: Development of LIS education oriented to e-science service requirements, de Si, Zhuang, Xing e Guo (2013), é o mais antigo de todos os documentos obtidos dentre os 34 identificados, e, portanto, tenderia a ser o mais citado, o que ficou constatado, com 18 citações. O artigo focaliza as qualificações dos profissionais que buscavam ser contratados em portais de emprego na China associadas aos termos "scientific data management", "data service", "data curation", "e-Science", "e-Research" e "data specialist", e buscou apresentar seus resultados de forma útil ao desenho de novos currículos de formação de profissionais das *iSchools* aplicados ao campo de "scientific data management" e "e-science" no contexto daquele país. As habilidades, as qualificações, os deveres e as responsabilidades identificados como mais valorizados na contração de profissionais para o "scientific data management" e "e-science" foram (SI; ZHUANG; XING; GUO, 2013): a) capacidade para trabalho em equipe; b) boa comunicação; c) relacionamento interpessoal; d) uso de ferramentas de curadoria de dados; e) padrões de metadados.

Si, Zhuang, Xing e Guo (2013) não analisaram demandas relacionadas a outros tipos de profissionais ligados à ciência de dados, como nas áreas de "big data analyst" ou mesmo "data scientist".

Educating Data Management Professionals: A Content Analysis of Job Descriptions, de Chen e Zhang (2017), apresenta análise de qualificações desejáveis requeridas junto a profissionais de "library and information science" no início de 2015, nos EUA, voltados à função de "librarian data management". Os resultados da pesquisa apontaram que "os candidatos mais bem sucedidos deveriam estar aptos a apoiar professores e alunos na coleção, gerenciamento e análise de dados de pesquisa". Não foram analisadas por Chen e Zhang (2017), quaisquer demandas relacionadas a outros tipos de profissionais ligados à ciência de dados, como nas áreas de "big data analyst" ou mesmo "data scientist".

Research Synthesis Methods and Library and Information Science: Shared Problems, Limited Diffusion, de Sheble (2016), buscou identificar, por meio de técnicas de pesquisa bibliográfica, as intersecções de temas de pesquisa que relacionam a "library and information science" com o uso dos métodos de RSM (*Research Synthesis Methods*), envolvendo a Revisão Sistemática e a Meta-análise. Os métodos de RSM são fortemente usados em pesquisa biomédica para a síntese de evidências. Os principais tópicos identificados na interseção entre "library and information science" e RSM foram: "open access", "information retrieval", "bias and research information ethics", "referencing practices", "citation patterns" e "data science". A aplicabilidade de "data science" foi evidenciada em situações nas quais algoritmos de aprendizagem de máquina poderiam auxiliar na identificação automática das fontes de informação mais relevantes para leitura detalhada (SHEBLE, 2016), o que exemplifica o uso de *machine learning* para finalidades de classificação.

Reproducible research and GIScience: an evaluation using AGILE conference papers, de Nuest *et al.* (2018), desenvolve um conjunto de critérios para a produção de pesquisa reproduzível, fundamentado na plena documentação, disponibilização e permanência de acesso aos dados, métodos e critérios para produção de resultados em artigos científicos, com foco na produção científica na área de *GIS Science* (Sistemas de Informação Geográficos). Foram então exploradas, junto aos autores de artigos de conferências na área de *GIS Science*, as razões pelas quais seus artigos apresentaram falhas no atendimento aos critérios de reprodutibilidade já reconhecidos. Dentre as cinco razões indicadas pelos autores, as mais relevantes para o cometimento das falhas foram, em ordem decrescente de relevância: (1) restrições legais; (2) falta de tempo; (3) falta de ferramentas; (4) falta de incentivos; e (5) falta de conhecimento.

Geography and computers: Past, present, and future, de Arribas-Bel e Reades (2018), explora de que forma a ciência de dados pode conduzir os estudos quantitativos em geografia para a criação de uma *Geographic Data Science*, que seria uma forma de ciência de dados consciente da espacialidade do dado geográfico, inclusive nas dimensões de crítica social e humanista presentes nos estudos geográficos. Segundo os autores, o desenvolvimento de uma *Geographic Data Science* combinaria a "tradição de *pensamento espacial* prevalente na geografia computacional e em sistemas de informação geográficos, com abordagens modernas para captura, transformação, processamento e análise evidentes na ciência de dados" (ARRIBAS-BEL; READES, 2018, p. 6), sem desconsiderar as questões epistemológicas, metodológicas e mesmo políticas que pautam o desenvolvimento histórico da geografia.

Data Science as a new frontier for Design, de Kazakci (2015), explora como as teorias do design – *design science*, *design research*, *design thinking* (BROWN, 2009; BIRKHOFER, 2011) – podem explicar as contribuições da ciência de dados para o aprimoramento do avanço científico na chamada *e-science*, ciência baseada em análise intensiva de dados. Para tal, o autor descreveu o caso chamado de *HiggsML Challenge* (Desafio de aprendizagem de máquina Higgs), no qual o desenvolvimento de soluções colaborativas de algoritmos de aprendizagem de máquina para análise dos dados gerados por um acelerador de partículas levou ao aprimoramento da compreensão do Bóson de Higgs. O caso envolveu 1.785 times de todo o mundo em uma espécie de evento do tipo *hackaton* e foi considerado muito bem-sucedido, inclusive por envolver pessoas com muito pouco conhecimento de física na produção de resultados relevantes.

A abertura de acesso aberto aos dados científicos *on-line*, o uso de uma plataforma *on-line* (Kaggle) para "gamificação" de todo o processo de formação dos times, o uso de repositórios de código aberto de programas de computador compartilhado em plataformas como o GITHub, a troca de mensagens em fóruns *on-line* e a exploração simultânea de várias estratégias distintas para análise de dados, algumas delas não convencionais, foram algumas das práticas marcantes que caracterizam os processos de trabalho dos cientistas de dados durante o *HiggsML Challenge*.

Knowledge mobilization for community resilience: perspectives from data, informatics, and information science, de Virapongse, Duerr e Metcalf (2019), apresenta resultado de uma discussão mediada com uma comunidade nos EUA que busca desenvolver um arcabouço (*framework*) de dados, métodos, ferramentas e processos que objetivam aprimorar a resiliência de comunidades frente à ocorrência de desastres naturais ou artificiais. A discussão visava identificar as principais recomendações para mobilização de informação no processo de aprimoramento da resiliência de comunidades. Dentre elas, foram elencadas e detalhadas:

a) Esclarecer o conceito de resiliência de comunidades, a fim de permitir identificar as necessidades de dados (fatores de riscos de natureza diversa, especialmente ambientais);

b) Priorizar suporte institucional para a resiliência de comunidades;

c) Garantir a acessibilidade e usabilidade aos dados;

d) Preencher os vazios no ciclo da informação, com vistas a promover a interação entre os provedores de dados e os praticantes de resiliência em comunidades.

ANÁLISE E DISCUSSÃO

A breve exploração da atual condição da ciência de dados, sob a perspectiva bibliográfica, combinada com a exploração dos nove artigos que evidenciam algumas das interlocuções entre a ciência de dados e a ciência da informação mostra que a ciência de dados é uma área do conhecimento predominantemente envolvida com o desenvolvimento de inovações tecnológicas e com a produção de conhecimento científico de natureza quantitativa em um contexto de massiva produção de dados.

A ciência de dados é uma área aplicada, que se apropria de ferramental estatístico e computacional recentemente produzido e disponibilizado em repositórios de software aberto, como o GITHub, para produzir, em curto espaço de tempo, análises e sínteses, especialmente as de natureza preditiva, baseadas em dados de variam amplamente em volume, variedade e velocidade de geração.

As análises e sínteses produzidas de forma algorítmica pela ciência de dados geram elevado impacto sobre a decisão em negócios empresariais e sobre a compreensão de fenômenos científicos em todas as áreas do conhecimento humano.

A perspectiva epistemológica dominante na ciência de dados é positivista, quantitativa, social e historicamente ingênua, além de fundamentada no compartilhamento de ferramental coletivamente gerado e na participação em jogos sociais competitivos, isto é, "gamificados".

No que se refere às implicações das decisões geradas pelas previsões e classificações de eventos feitas por modelos desenvolvidos em ciência dos dados, pairam dúvidas sobre a validade do que é produzido, sobretudo pela ausência de explicabilidade e rastreabilidade dos processos dedutivos e indutivos empregados em algumas situações.

A breve exploração da co-ocorrência de palavras-chave na interlocução entre a ciência de dados e a ciência da informação evidencia o desafio educacional que a ciência da informação enfrenta para, de um lado, incorporar à ciência de dados os valores e responsabilidades do profissional da informação e, de outro lado, gerar novo valor em ciência da informação a partir da oportunidade produzida pelos *big data*, com a incorporação de ferramentas da inteligência artificial e da aprendizagem de máquina. Nesse sentido, quais princípios, tendências, designs de sistemas e riscos surgem na transformação da ciência da informação por princípios epistemológicos, modelos, métodos, ferramentas e valores da ciência de dados?

No que se refere às mensagens latentes nos nove artigos mais acoplados bibliograficamente na relação entre a ciência da informação e a ciência dos dados, cabem algumas considerações finais:

1. **A transformação paradigmática da ciência da informação** pela emergência dos *big data* é iminente e, para que essa transformação se processe em um sentido positivo, deve ser garantida a incorporação de valores e princípios da ciência da informação, tais como, segundo Wang (2018): a concepção de dados nos sentidos históricos e pragmáticos; o cuidado com a precisão, acurácia e credibilidade de dados; os processos e funções clássicos da biblioteconomia; além das teorias da documentação.

2. **Existem distintas formas de neutralidade na ciência da informação e na ciência de dados, e cada uma apresenta aspectos positivos e negativos**. Por um lado, a ciência de dados tende a assumir uma condição de neutralidade perante dados, informações e conhecimentos à sua disposição, sejam ou não esses dados validados do ponto de vista científico. Nesse sentido, a ciência de dados promove ativamente uma postura inovadora, aderente ao *design thinking* (VIRAPONGSE; DUERR; METCALF, 2019), criando novos modelos de natureza estatística e algorítmica, produzidos para domínios de conhecimento específicos, visando à imediata geração de novo conhecimento acionável para a tomada de decisão, baseada nos recursos à disposição. Por outro lado, a ciência da informação clássica se promove como mediadora das necessidades de informação de seus usuários (ZINS, 2007) frente à organização dos dados, de informações e conhecimentos científicos já produzidos, validados e registrados, com os quais essa ciência também se apresenta neutra, sem assumir caráter julgador. Entretanto, as práticas da ciência de dados não se apresentam como ativas promotoras de intervenções, em especial no processo decisório.

3. **As escolas de ciência da informação devem se movimentar com celeridade compatível com a avalanche tecnológica dos big data** (ORTIZ-REPISO; GREENBERG; CALZADA-PRADO, 2018; VIRAPONGSE; DUERR; METCALF, 2019; WANG, 2018) na produção de novos currículos para oportunizar a sobrevivência e o reconhecimento dos profissionais da informação, ou mesmo a preservação dos valores profissionais e éticos em um contexto de geração de novas profissões.

4. **Uma potencial dominância dos conceitos empregados na geografia durante a interlocução entre a ciência de dados e a ciência da informação merece exploração futura**, como sugerem os estudos de Nuest *et al.* (2018) e Arribas-Bel e Reades (2018). Ainda, com base em Nuest *et al.* (2018), o conjunto das razões pelas quais as pesquisas científicas em determinada área de conhecimento apresentam problemas com reprodutibilidade sugerem aspectos que podem ser considerados relevantes em contribuições da ciência da informação para o aprimoramento da ciência de dados e estariam relacionados ao desenvolvimento de epistemologias, metodologias, instrumentos legais, códigos de conduta profissional, incentivos mercadológicos e ferramentais críticos para aprimoramento da prática dos cientistas de dados, no sentido de melhor compreensão das dimensões cognitivas e sociais envolvidas com os dados, métodos e critérios que empregam na produção de suas análises.

5. **As contribuições da ciência de dados para a evolução das demais áreas do conhecimento científico e suas correspondentes profissões são marcantes e distintas, sugerindo a possibilidade de que a ciência de dados se constitua como um novo paradigma científico**. Se os erros e acertos decorrentes da aplicação massiva dos métodos e técnicas da ciência de dados nas empresas para a criação de soluções de *data analytics* apresentam problemas éticos que devem ser endereçados com aprimoramentos legais, normativos e profissionais, a ciência de dados evidencia a oportunidade da emergência da *e-science*. Essa nova forma de fazer ciência é caracterizada no estudo de Kazakci (2015).

6. **O problema da resiliência das comunidades vem a ser um arquétipo dos problemas enfrentados pela humanidade no atual momento em que nos encontramos, de pandemia causada pelo SARS-Cov-2**. Nesse sentido, o avanço da ciência cidadã é fundamentalmente baseado nas ferramentas da ciência de dados e depende do preenchimento dos imensos vazios no ciclo da informação, que foram levantados por Virapongse, Duerr e Metcalf (2019), no contexto da promoção da resiliência em comunidades, como:

- Organizar oficinas de trabalho e tutoriais para auxiliar os praticantes de análise e síntese de dados em comunidades a acessarem dados e a usarem ferramentas para a sua análise;
- Ensinar habilidades de colaboração, facilitação, liderança, gerenciamento e trabalho em equipe para produtores de dados;
- Criar papéis ativos e distintivos para pessoas atuantes no ciclo da informação e do conhecimento, tais como intermediários no acesso aos dados, intérpretes de dados científicos, mediadores, promotores de engajamento em comunidades e contadores de histórias;
- Ligar praticantes de análise e síntese de dados em comunidades às redes onde se encontram os produtores de dados;
- Identificar comunidades que precocemente adotam inovações, para intensificar colaborações e parcerias, promovendo exemplos de usos de dados e ferramentas junto às comunidades;
- Promover encontros para criação de projetos pilotos envolvendo produtores de dados e praticantes de análise e síntese de dados em comunidades;
- Criar fóruns para encontros regulares de distintas partes interessadas no ciclo da informação relativa à resiliência de comunidades;
- Promover organizações para mediação da comunicação entre produtores de dados e praticantes de análise e síntese de dados em comunidades.

As tradições da ciência da informação presentes na releitura dos conceitos de bibliotecas são extremamente pertinentes no preenchimento desses vazios.

CONCLUSÕES

Este artigo buscou situar, do ponto de vista bibliográfico, o atual estado da ciência dos dados em sua relação com a ciência da informação. Foram empregadas pesquisas em uma base de dados bibliográfica, obtendo-se registros que foram analisados de forma visual, usando as técnicas de análise de co-ocorrência e acoplamento bibliográfico. As perspectivas de um breve entrelaçamento entre a ciência de dados e a ciência da informação são traçadas em ambas as direções, apontando não apenas para uma oportunidade de mútuo benefício das áreas e de seus praticantes, mas para uma necessidade de que isso ocorra em benefício de uma sociedade mais resiliente, sustentável, democrática e ética.

REFERÊNCIAS

ARRIBAS-BEL, D.; READES, J. Geography and computers: Past, present, and future. *Geography Compass*, [*s.l.*], v. 12, n. 10, p. e12403, Sep. 2018. Disponível em: https://doi.org/10.1111/gec3.12403. Acesso em: abr. 2021.

BIRKHOFER, H. (ed.). *The Future of Design Methodology*. London: Springer-Verlag, 2011.

BROWN, T. *Change by Design:* how design thinking transforms organizations and inspires innovation. USA: Harper Collins, 2009.

BUCKLAND, M. The physical, mental and social dimensions of documents. *Proceedings from the Document Academy*, Ohio, USA, v. 3, n. 1, p. 1-6, 2016. DOI https://doi.org/10.35492/docam/3/1/4. Disponível em: https://ideaexchange.uakron.edu/docam/vol3/iss1/4. Acesso em: maio 2021.

CHEN, H.; ZHANG, Y. Educating Data Management Professionals: a content analysis of job descriptions. *Journal of Academic Librarianship*, [USA], v. 43, n. 1, p. 18–24, Jan. 2017. Disponível em: https://doi.org/10.1016/j.acalib.2016.11.002. Acesso em: maio 2021.

DAVENPORT, T. H.; PATIL, D. J. Data Scientist: the sexiest job of the 21st Century. *Harvard Business Review*, [*s.l.*], v. 90, n. 10, p. 70–76, Oct. 2012. Disponível em: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century. Acesso em: maio 2021.

KAZAKCI, A. O. Data science as a new frontier for design. *In*: INTERNATIONAL CONFERENCE ON ENGINEERING DESIG, 15., 2015, Milan, Italy. *Proceedings [...]*.Glasgow: Design Soc, 2015. Disponível em: https://www.designsociety.org/publication/37969/

Modelo populacional para análise de genealogia acadêmica: evidências sobre crescimento acadêmico no Brasil

Rafael Jeferson Pezzuto Damaceno
Doutor em Ciência da Computação pela Universidade Federal do ABC (UFABC) - Santo André, SP - Brasil.
http://lattes.cnpq.br/2552938367742788
E-mail: rafael.damaceno@ufabc.edu.br

Maximiliano Barbosa da Silva
Doutor em Economia pela Universidade de São Paulo (USP) – SP - Brasil. Professor da Universidade Federal do ABC (UFABC) - Santo André, SP - Brasil.
http://lattes.cnpq.br/2344017108313395
E-mail: maximiliano.silva@ufabc.edu.br

Jesús Pascual Mena Chalco
Pós-Doutorado pela Universidade de São Paulo (USP) – SP - Brasil. Doutor em Ciências da Computação pela Universidade de São Paulo (USP) – Brasil, com período co-tutela em Instituto Nacional de Matemática Pura e Aplicada (IMPA) – Brasil. Professor pela Universidade Federal do ABC (UFABC) - Santo André, SP - Brasil.
http://lattes.cnpq.br/4727357182510680
E-mail: jesus.mena@ufabc.edu.br

Submetido em: 26/09/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021.

RESUMO

Estudos recentes têm analisado a formação de novos cientistas por meio das relações de orientação em nível de pós-graduação. No entanto, essa literatura é eminentemente estática, no sentido de que não se aprofunda na atuação dos acadêmicos no decurso de suas carreiras. A fim de contribuir para o preenchimento dessa lacuna, este trabalho expande a análise das relações de orientação para um modelo populacional de crescimento, que contabiliza, anualmente, medidas a elas relacionadas. O modelo populacional de crescimento é aplicado em um conjunto composto por mais de 1 milhão de relações formais de orientação nos níveis de mestrado e doutorado, além de supervisões de pós-doutorado. As três principais contribuições deste trabalho correspondem à: (a) construção de um modelo para analisar o crescimento de grafos de genealogia acadêmica; (b) identificação da evidência de decréscimo, nos últimos anos, do percentual de acadêmicos que se tornam professores orientadores; e (c) identificação da evidência de maior produtividade observada nos professores seniores em comparação com outros atores.

Palavras-chave: Genealogia acadêmica. Carreira acadêmica. Modelo populacional de crescimento.

Population model for analyzing academic genealogy: evidence on growth academic in Brazil

ABSTRACT

Recent studies have analyzed the formation of new scientists at the master's and doctorate levels. However, such analyzes does not take into account how the academics' career are. In this sense, this work expands the analysis of the relationships established between advisors and advisees towards a population growth model. We apply this model to a group composed of more than 1 million formal mentoring relationships established at the masters and doctoral levels, and at the post-doctoral supervision. The main contributions of this work are as follows. (a) The building of a population model applicable to academic genealogy graphs, (b) the indication of a decrease in the percentage of academics who become advisors themselves, and (c) the indication that senior academics have higher productivity when compared to other academic categories.

***Keywords**: Advisor-advisee relationships. Academic carrer. Population model.*

Modelo poblacional para analizar genealogía académica: evidencia sobre el crecimiento académico en Brasil

RESUMEN

Recientes estudios analizan la formación de nuevos científicos por medio de las relaciones de orientación a un nivel de post-graduación. Sin embargo, esa literatura es eminentemente estática en el sentido de que no profundiza en la actuación de académicos a largo de sus carreras. Para contribuir con la disminución de esa laguna este trabajo expande el análisis de las relaciones de orientación para un modelo de crecimiento poblacional que contabiliza, anualmente, diferentes medidas relacionadas a ellas. El modelo de crecimiento poblacional es aplicado en un conjunto de datos compuesto por más de 1 millón de relaciones formales de orientaciones en los niveles de maestria y doctorado, además de las supervisiones de post-doctorado. Las tres principales contribuciones de este trabajo corresponden a: (a) la construcción de un modelo para analizar el crecimiento de grafos de genealogía académica, (b) la indicación de una disminución en el porcentaje de académicos que se convierten ellos mismos en asesores, y (c) la indicación de que los académicos seniors tienen una mayor productividad en comparación con otras categorías académicas.

***Palabras clave**: Genealogía académica. Carrera académica. Modelo de población.*

INTRODUÇÃO

Modelo populacional é uma ferramenta matemática utilizada para analisar como populações (e.g., indivíduos, bactérias, entre outros) crescem. Alguns modelos eminentes são o de Thomas Malthus, no qual a taxa de crescimento é proporcional à totalidade da população existente (MALTHUS, 1992), e o de Pierre François Verhulst, em que a taxa de crescimento é limitada a uma capacidade máxima de população (BACAËR, 2011).

Tais modelos têm sido adaptados e aplicados a redes de coautoria e, um exemplo, é o estudo de Wu *et al.* (2018), que desenvolvem um sistema populacional a partir de publicações da área de Ciência da Computação registradas na *Digital Bibliography & Library Project* (DBLP). O sistema por eles proposto, além de contabilizar a população de autores anualmente, diferencia os pesquisadores conforme e com quem colaboram na produção de artigos.

De acordo com os autores, nos últimos anos, tem havido um aumento no número de pesquisadores que surgem no modelo populacional por intermédio de autoria única ou coautoria com pesquisadores também recém-chegados no sistema.

Neste trabalho, propõe-se **um novo modelo populacional,** usando, como base, dados advindos das relações formais de orientação (genealogia acadêmica). Esse **modelo foi aplicado em um conjunto abrangente da genealogia acadêmica do Brasil,** composta de mais de 1 milhão de pessoas e relações.

O trabalho permite evidenciar como as relações de orientação (mestrado e doutorado) e supervisão (pós-doutorado) acadêmicas têm progredido no decorrer de dada temporalidade e que papéis os professores orientadores e alunos orientados/supervisionados, isto é, seus descendentes, têm exercido nessa progressão. Em concreto, com esse modelo populacional e os dados nacionais, seis questões de pesquisa são respondidas:

- QP1:Como é a entrada (surgimento) de professores e descendentes ao longo dos anos?
- QP2: Como é a saída (cessamento) de professores e descendentes ao longo dos anos?
- QP3:Qual é a população de professores e descendentes no Brasil?
- QP4:Quantos professores estão deixando de orientar/supervisionar ao longo dos anos?
- QP5:Quantos descendentes estão se tornando professores ao longo dos anos?
- QP6: Há diferenças na quantidade e no tipo de publicações, considerando os papéis (professor ou descendente) exercidos por acadêmicos?

O restante do trabalho está organizado da seguinte forma: na seção seguinte, são apresentados os conceitos relacionados ao modelo populacional de crescimento; na seção Conjunto de Dados, são descritos os dados utilizados neste trabalho; na seção "Resultados", são respondidas questões acerca da aplicação do modelo no grafo de genealogia acadêmica do Brasil; por fim, a última seção conclui a presente investigação e apresenta possíveis trabalhos futuros.

MODELO POPULACIONAL DE GENEALOGIA ACADÊMICA

O insumo para geração do modelo populacional de genealogia acadêmica é uma estrutura hierárquica que apresenta as relações formais de orientação entre professores e alunos em dada cronologia. Em outras palavras, precisa-se de um grafo direcionado, com informação temporal nas arestas, que apontam no sentido orientador-aluno.

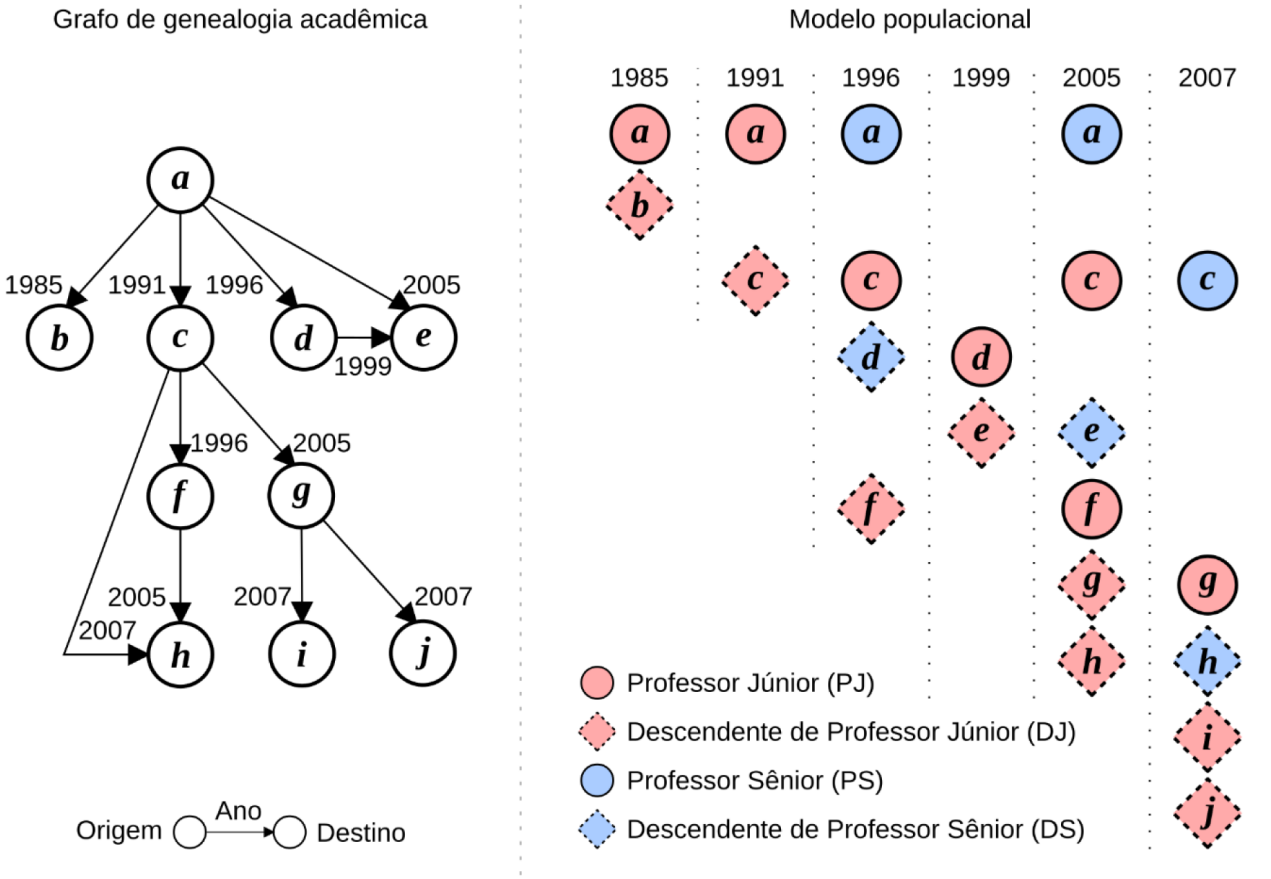
No modelo, os acadêmicos são divididos em quatro tipos, conforme os papéis que exercem durante sua carreira. Um acadêmico é professor orientador júnior (PJ), se atua como orientador há menos de anos[1]; é professor orientador sênior (PS), se atua como orientador há ou mais anos; é descendente júnior (DJ), se foi orientado por PJ e ainda não orientou; é descendente sênior (DS), se foi orientado pela última vez por PS e ainda não orientou.

Diferenciar os acadêmicos nesses quatro tipos permite, por exemplo, verificar como atuam, em termos de orientação ou produção bibliométrica, os alunos orientados por professores mais experientes (PS) em comparação com aqueles orientados por professores menos experientes (PJ). Pode-se também estudar como é a produtividade de PS em relação à de PJ, do ponto de vista de medidas bibliométricas e de formação acadêmica.

[1] O valor da variável reflete o tempo mínimo de experiência que um acadêmico precisa ter para ser considerado sênior, de modo que, quanto maior for o seu valor, maior será a exigência para considerá-lo como sênior. O modelo não se restringe, no entanto, a essa definição; outras características podem ser associadas a senioridade, tais como: o número de alunos orientados, o ano da primeira publicação realizada, entre outras.

A figura 1 ilustra o nosso modelo populacional como um grafo de genealogia acadêmica, destacando as quatro subpopulações definidas acima. A figura também permite observar como ocorre a migração de um indivíduo de um tipo de acadêmico para outro considerando. Especificamente, à esquerda está representado o grafo de genealogia acadêmica em que cada vértice/letra representa um acadêmico e cada aresta representa a orientação/supervisão realizada por um professor a um aluno no ano indicado. Esse exemplo é traduzido para o modelo populacional representado à direita.

Figura 1 – Exemplo de transformação de grafo de genealogia em indicadores considerados no Modelo Populacional



Fonte: Elaborado pelos autores (2020).

Nesse modelo populacional, os acadêmicos possuem um ciclo de vida, que pode envolver a transição de descendente júnior ou sênior para professor júnior e, então, para professor sênior. Denota-se como produtividade (P) o modo como ele atua no decurso desses estágios, do ponto de vista de medidas bibliométricas. Precisamente, P é uma medida para contabilizar o de número de artigos, capítulos de livros e livros publicados por ano, por tipo de acadêmico.

Um acadêmico inicia seu ciclo de vida ao entrar no sistema populacional, o que pode ocorrer de duas formas: (i) ser orientado pela primeira vez ou (ii) orientar pela primeira vez. No primeiro caso, na situação em que é orientado por PJ, assume o papel de DJ - ou, analogamente, ao ser orientado por PS, assume o papel de DS. No segundo caso, ao orientar pela primeira vez, assume o papel de PJ – como não estava na população até aquele momento, é considerado um caso de partenogênese, isto é, um nascimento sem orientador. No exemplo, o acadêmico orienta, em 1985, seu primeiro aluno, ; nesse caso, tanto quanto entram no sistema em 1985, na condição de PJ e na condição de DJ.

Conforme recebem e oferecem orientação, os acadêmicos transitam para outros tipos populacionais. No caso de um DJ ou DS, ao orientarem pela primeira vez, transitam para o tipo PJ. Essa é a situação em que um ex-aluno passa a atuar como professor. Ilustração disso é o acadêmico , que foi orientado em 1991, entrando no sistema como DJ, e orientou seu primeiro aluno cinco anos depois, transitando para o papel de professor júnior, PJ.

Outra situação de transição é a de um orientador PJ que, ao possuir anos de experiência na atividade de orientação, assume o papel de PS. A partir desse momento, ao realizar novas orientações, gerará acadêmicos DS. Neste trabalho considera-se, isto é, um acadêmico PS é aquele que possui pelo menos dez anos desde a realização de sua primeira orientação. Como evidencia a exemplificação, o acadêmico passa a atuar como PS a partir de 1996, já que sua primeira orientação ocorreu em 1985.

A saída de acadêmicos do sistema está relacionada ao cessamento da atividade de orientação ou à não transição de tipos. Um acadêmico DS ou DJ deixa o sistema ao não transitar para PJ. Isso significa que, em termos de formação de recursos humanos, não prossegue na academia. No exemplo, o acadêmico , que recebeu duas orientações, uma em 1999 e outra em 2005, não atua como professor no tempo futuro. Um acadêmico PJ ou PS deixa o sistema ao não realizar mais orientações, o que é contabilizado do ponto de vista dos anos subsequentes à sua última orientação. Para isso, considera-se uma janela temporal de anos desde o último ano da série analisada. Como mostra o exemplo, o acadêmico realiza sua última orientação em 2005, porém, para que seja possível fazer essa afirmação, é preciso ter dados de referência até anos.

No contexto de saída de acadêmicos, pode-se definir inatividade de orientação tanto para professores, quanto para descendentes. No primeiro caso, contabilizam-se os acadêmicos que deixaram de orientar após orientarem seu primeiro descendente. E no segundo, contabilizam-se os descendentes que não transitaram para o papel de orientador. Naturalmente, a caracterização da inatividade de orientação depende do horizonte temporal futuro utilizado. Para acadêmicos ingressantes em 2008, a título de ilustração, são considerados como cessantes, isto é, que saíram, aqueles que não realizaram orientações ou não transitaram de tipo no período de 2009 até o último ano disponível de informações.

CONJUNTO DE DADOS

Este estudo utiliza dados de genealogia acadêmica do Brasil, obtidos de método criado no trabalho de Damaceno *et al.* (2019). Trata-se de um grafo composto por 1.272.590 vértices (acadêmicos mestres e/ou doutores) e 1.404.109 arestas (relações formais de orientação de mestrado e doutorado, e supervisões de pós-doutorado). Desses vértices, 502.307 são doutores[2] e 770.283 são mestres.

[2] Um acadêmico é considerado doutor, no grafo, se (a) orientou/supervisionou em qualquer nível, (b) foi orientado em Doutorado ou (c) foi supervisionado em pós-doutorado.

Pode-se associar uma série de medidas topológicas a cada vértice de um grafo. As medidas mais relacionadas a este trabalho são: graus de entrada, saída e total, número de primos, número de irmãos e índice genealógico. No contexto deste trabalho, grau de entrada é o número de orientadores que um acadêmico possui, grau de saída é o número de alunos que um acadêmico orientou e grau total é a soma dos graus de entrada e saída.

Número de primos e de irmãos são medidas análogas às de genealogia familiar. O índice genealógico de um acadêmico é o maior número de seus orientados, que possuem, no mínimo, orientados cada um (ROSSI *et al.*, 2017). Essas medidas topológicas foram calculadas para o grafo e para os dados anuais do modelo populacional.

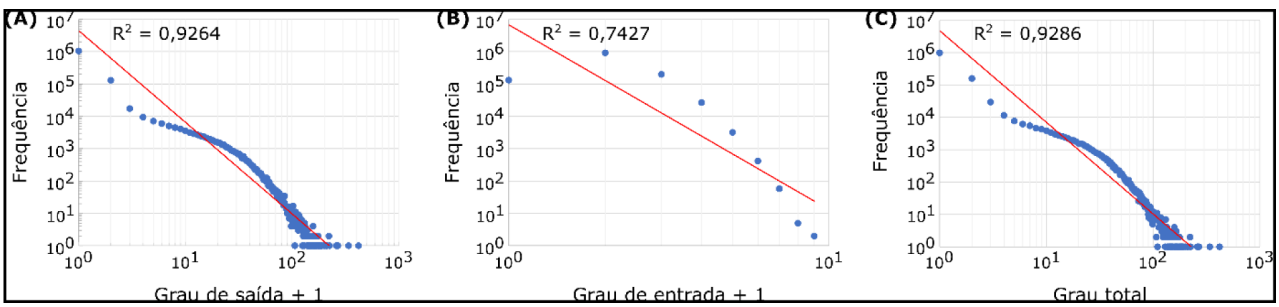
Frequentemente, verifica-se que a distribuição de grau dos vértices de uma rede social pode ser bem aproximada por uma lei de potência (PRICE, 1963; ALBERT, JEONG; BARABÁSI, 1999). Esse tipo de distribuição pode ser gerado, por exemplo, pelo modelo de ligação preferencial, em que um novo vértice nasce a cada período conectado, inicialmente, a certo número de outros vértices e o seu grau cresce, ao longo do tempo, proporcionalmente ao tamanho da sua vizinhança.

Uma propriedade das leis de potência é que elas possuem caudas mais pesadas do aquelas de uma distribuição Poisson, que resulta do modelo de formação de rede de Erdös e Rényi (1961), em que a incidência dos links sobre os vértices é determinada aleatoriamente.

Na medida em que diferentes modelos de formação de redes possuem implicações sobre a distribuição de grau dos vértices no grafo, é instrutivo comparar a distribuição empírica com aquela prevista pelos modelos teóricos, no sentido de iluminar o mecanismo pelo qual a rede social observada é formada. Em particular, uma lei de potência, que é gerada pelo modelo de ligação preferencial, transforma-se em uma reta quando representada, no plano cartesiano, em escala log-log. Por outro lado, a distribuição de grau gerada pelo modelo de Erdös e Rényi (1961) é estritamente côncava nesse espaço.

Por esse motivo, ou seja, a fim de ganhar uma sensibilidade quanto aos mecanismos de formação da rede genealógica da academia brasileira, na figura 2 apresenta-se a distribuição de grau de saída (A), de entrada (B) e do grafo não direcionado (grau total, em C) correspondente na escala log-log. Nos gráficos, também é apresentada a reta obtida pela regressão linear simples de . Essa reta corresponde à lei de potência que melhor se ajusta aos dados.

Figura 2 − Distribuição de grau do grafo de genealogia acadêmica



Fonte: Elaborado pelos autores (2020).

Como se pode perceber, as distribuições empíricas apresentam curvatura no plano log-log, sugerindo algum grau de aleatoriedade na formação das relações de orientação. Por outro lado, os coeficientes de ajustamento (R^2) são bastante elevados, superiores a 74%, indicando que o mecanismo de ligação preferencial pode ser um ingrediente relevante na formação das relações de orientação acadêmica (JACKSON; ROGERS, 2007).

MODELO POPULACIONAL APLICADO À ACADEMIA BRASILEIRA

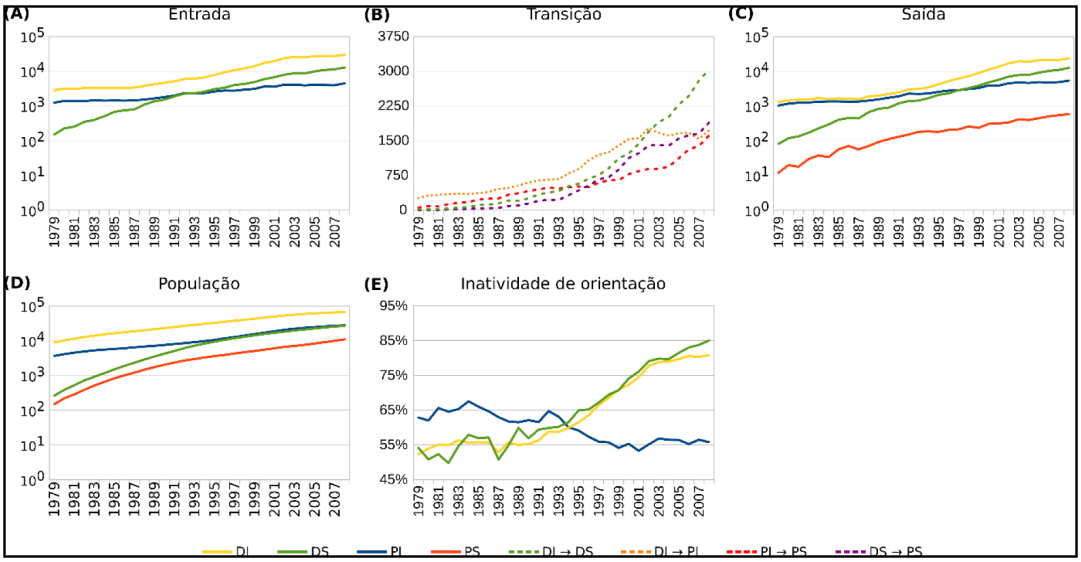
O modelo populacional foi aplicado para o contexto nacional mais abrangente. Acredita-se que esse tratamento traz uma visão inédita sobre a formação de acadêmicos no Brasil ao longo dos tempos. Este trabalho discute seis questões de pesquisa com o objetivo de verificar como a rede formada por esses acadêmicos se configura em relação ao modelo populacional de crescimento proposto.

Para isso, foram calculadas as entradas, transições e saídas, considerando-se uma janela temporal de 1900 até 2018, e são apresentados os resultados referentes ao período de 1979 até 2008 (30 anos). As três subseções expostas a seguir abordam as seguintes temáticas: (i) população de professores e descendentes; (ii) inatividade de orientação; e (iii) produtividade ponderada de professores e descendentes.

POPULAÇÃO DE PROFESSORES E DESCENDENTES

Esta subseção analisa a população de professores e descendentes que o Brasil está gerando no decorrer dos anos por tipo. Em 1979, o modelo possui uma população composta de 8.863 PJ, 195 PS, 17.467 DJ e 509 DS, o que totaliza 27.034 acadêmicos. No ano de 2008, essa população é de 460.244 acadêmicos, distribuídos em 43.375 PJ, 15.201 PS, 299.543 DJ e 102.125 DS. A figura 3 apresenta os dados referentes à entrada (A), transição (B), saída (C), população (D) e inatividade de orientação (E) para acadêmicos no período de 1979 até 2008.

Figura 3 – Entrada, transição, saída, população e inatividade de orientação para acadêmicos no período de 1979 até 2008



Fonte: Elaborado pelos autores (2020).

QP1: Como é a entrada (surgimento) de professores e descendentes ao longo dos anos?

Como pode ser observado na figura 3.A, o número de acadêmicos que entram no sistema tem crescido na cronologia investigada. Esses dados refletem, pelo menos em parte, o aumento do número de programas de pós-graduação que ocorreu no Brasil no período de 1996 até 2014, que foi, em média, de 6,4% e 6,5%, por ano, para os programas de mestrado e doutorado, respectivamente (CGEE, 2016). Esse crescimento, entretanto, não é uniforme em todo o período.

De fato, analisando os dados do CGEE (2016), verifica-se que, em alguns anos, tais como 1999, 2000, 2006 e 2013, as taxas de crescimento, para os programas de doutorado, foram muito maiores que a média (8,20%, 9,18%, 8,02 e 13,74%, respectivamente). Para os programas de mestrado, os anos que obtiveram os picos de crescimento foram 1999 (8,21%), 2002 (8,42%), 2006 (9,70%) e 2011 (10,21%).

No início da década de 1980 a diferença entre o número de ingressantes DJ e DS é maior em cotejo com o final da década de 2000. Ainda, por volta de 1993, passa a ingressar no sistema maior quantidade de acadêmicos DJ, seguida de DS e PJ. No caso das transições, na figura 3.B, observa-se crescimento no decurso de todos os anos do período analisado; exceção é feita para as transições de DJ para PJ que atingiram um platô no período de 2002 até 2008.

Uma genealogia que apresenta maior quantidade de acadêmicos PS pode ser considerada mais madura. No entanto, o que se percebe na genealogia brasileira é que, sob a ótica de relações de mestrado, doutorado e supervisões de pós-doutorado, em termos absolutos, as futuras gerações de pesquisadores são preparadas por acadêmicos relativamente jovens, dado o maior número de DJ em relação a DS. Em termos relativos, o número de orientandos de um PJ é similar ao de um PS.

QP2: Como é a saída (cessamento) de professores e descendentes ao longo dos anos?

A figura 3.C apresenta a saída de acadêmicos do sistema populacional. Nota-se, em termos absolutos, que PS saem do sistema em menor quantidade quando comparados a PJ, o que também ocorre para DJ quando comparado a DS. Em parte, esse fato reflete a maior entrada de DJ no sistema.

Ainda, percebe-se que, em 2008, o número de acadêmicos PJ que deixam o sistema atinge o mesmo nível que o de acadêmicos DS. O número de acadêmicos DJ e DS que saem do sistema vêm aumentando de 1999 até 2008, ao compará-lo com o número de acadêmicos PJ e PS.

QP3: Qual é a população de professores e descendentes no Brasil?

Como esperado, a população de cada tipo de acadêmico cresce com o passar dos anos, já que há mais acadêmicos ingressando ou transitando do que saindo, como pode ser observado na figura 3.D. Especificamente em relação aos acadêmicos descendentes, por um lado, os seniores ocorrem em menor número em todos os anos. Por outro lado, os descendentes juniores ocorrem em maior número.

No que diz respeito aos professores, há mais juniores do que seniores em todos os anos analisados. Vê-se também que, por volta de 1995, o número de descendentes seniores passou a ser próximo ao de professores juniores.

INATIVIDADE DE ORIENTAÇÃO

Esta seção analisa a quantidade de professores que orientam apenas uma vez em todas as suas carreiras e a quantidade de descendentes que não passam a atuar como professores. Essa análise pode evidenciar indícios de abandono do meio acadêmico por parte de professores e descendentes. A figura 3.E apresenta os resultados referentes à inatividade de orientação, que permitem responder às duas perguntas indicadas a seguir.

QP4: Quantos professores estão deixando de orientar, ao longo dos anos?

Concernente à inatividade de orientação de professores juniores, no ano inicial analisado, essa taxa era de 62,5%, isto é, de todos os PJ que surgiram em 1979, esse é o percentual de docentes que não realizaram outras orientações no período de 1980 até 2008. Esse percentual atinge um valor mínimo em 1999 (aproximadamente 53%), e então passa crescer novamente, atingindo pouco mais de 55% em 2008. Esses resultados indicam que a maior parte dos professores juniores orienta apenas uma vez em nível de pós-graduação, portanto, não se transforma em professores seniores ou prossegue na carreira de professor orientador.

QP5: Quantos descendentes estão se tornando professores ao longo dos anos?

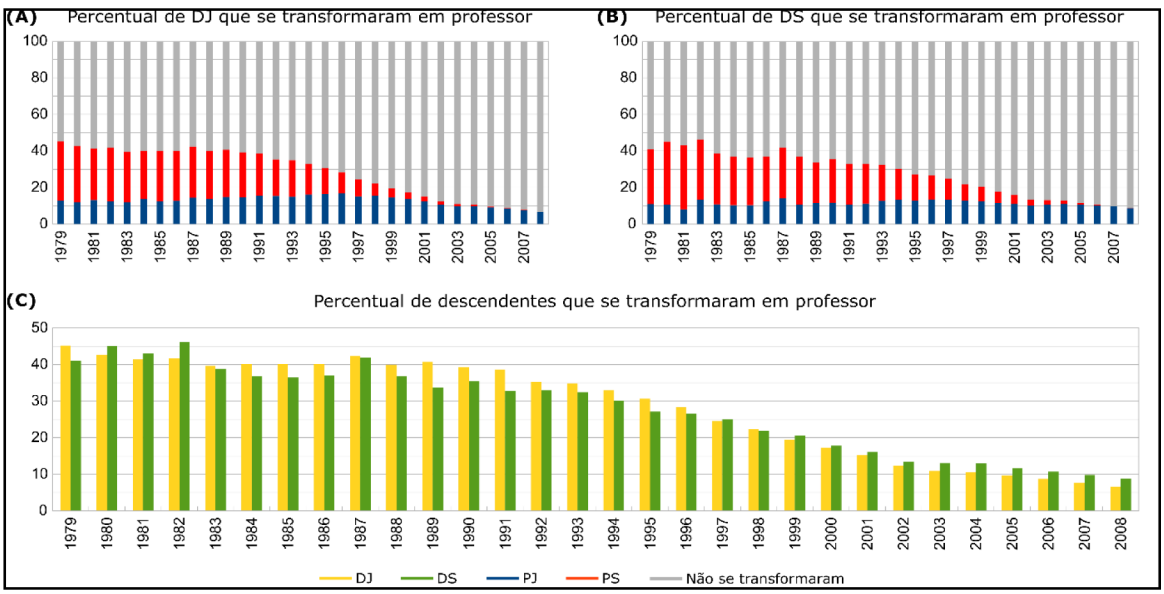
A figura 3.E também mostra a medida de inatividade de orientação para descendentes juniores e seniores, isto é, a população desses acadêmicos que não transitou para professor.

Durante o período analisado observa-se que, para ambos os tipos de descendentes, tem aumentado o valor percentual dessa medida, isto é, menos descendentes estão se transformando em professor.

Vê-se, no entanto, que, nos últimos anos do período, o percentual de inatividade de orientação de DS tem sido maior em relação ao de DJ. Ainda, a figura 3.E ilustra que ambos os tipos de descendentes têm deixado a academia em quantidade semelhante.

Na figura 4 notam-se os percentuais de descendentes juniores (A) e seniores (B) que vêm a se tornar professores juniores e seniores. Também, na figura 5.C, é apresentado o percentual total de descendentes que se transformam em professores, contabilizando-se os dois tipos de descendentes. As frequências foram calculadas partindo-se do ponto de vista do ano de origem dos descendentes, isto é, considerando que um descendente cujo ano de entrada foi t, verificou-se se esse descendente se tornou (ou não se tornou) professor (júnior ou sênior) após o ano *t*.

Figura 4 – Percentual de acadêmicos originários como DJ ou DS, que se transformaram em professor - período de 1979 até 2008



Fonte: Elaborado pelos autores (2020).

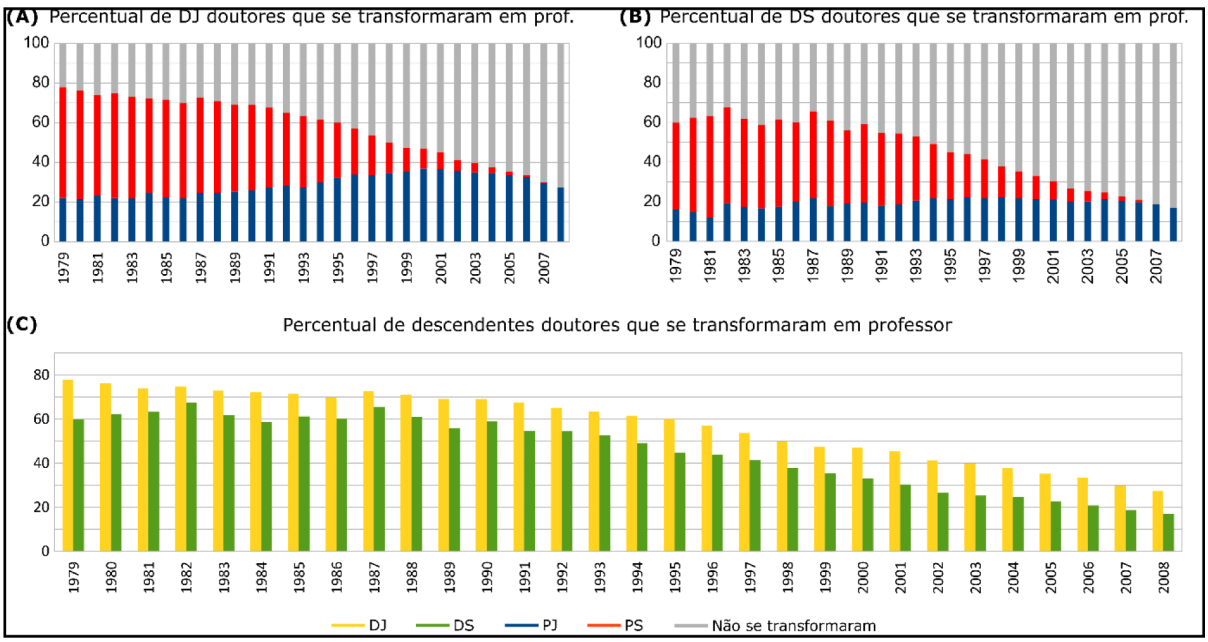
Os dados mostram que, no início da série, cerca de 45% dos acadêmicos DJ e 41% dos DS tornaram-se professor (em algum ano no futuro). No final da série, em 2008, cerca de 6,6% dos descendentes DJ e 8,7% dos descendentes DS transformaram-se em professor. De 1980 a 1982, em 1997 e a partir de 1999, maior proporção de descendentes juniores passa à categoria de professor, no futuro, em relação a descendentes juniores. Ainda, os gráficos destacam que, independentemente do tipo de orientação obtida (se de professor júnior ou de professor sênior), cada vez menos descendentes estão atuando como professor orientador nos anos seguintes analisados

Cabe destacar que essas transformações consideram tanto acadêmicos com nível de mestrado quanto com nível de doutorado.

Nessa situação, os dados podem indicar (a) uma preferência dos mestres em deixar a academia e (b) uma maior exigência para iniciar a carreira docente, dado que, nos últimos anos, as instituições de ensino têm exigido o nível de doutorado.

De modo mais específico, foi avaliado também o percentual de doutores que se converteram em professor orientador, permitindo verificar indícios de afastamento ou não da carreira acadêmica. Para isso, foi preciso restringir os dados de transições a apenas acadêmicos doutores. Nesse sentido, a **Figura 5** apresenta o percentual de acadêmicos DJ doutor e DS doutor que se transformaram em PJ ou PS.

Figura 5 – Percentual de acadêmicos doutores originários como DJ ou DS, que se transformaram em professor - no período de 1979 até 2008



Fonte: Elaborado pelos autores (2020).

Nesse caso, observa-se que, em todos os anos, um maior percentual de doutores com origem júnior se transformou em professores doutores, em relação àqueles com origem sênior. Nota-se também um padrão de decrescimento do percentual de descendentes que acessam a carreira docente. Mais doutores estão se formando, logo, as instituições, principalmente no Brasil, podem estar sendo incapazes de absorvê-los.

PRODUTIVIDADE PONDERADA DE ACADÊMICOS

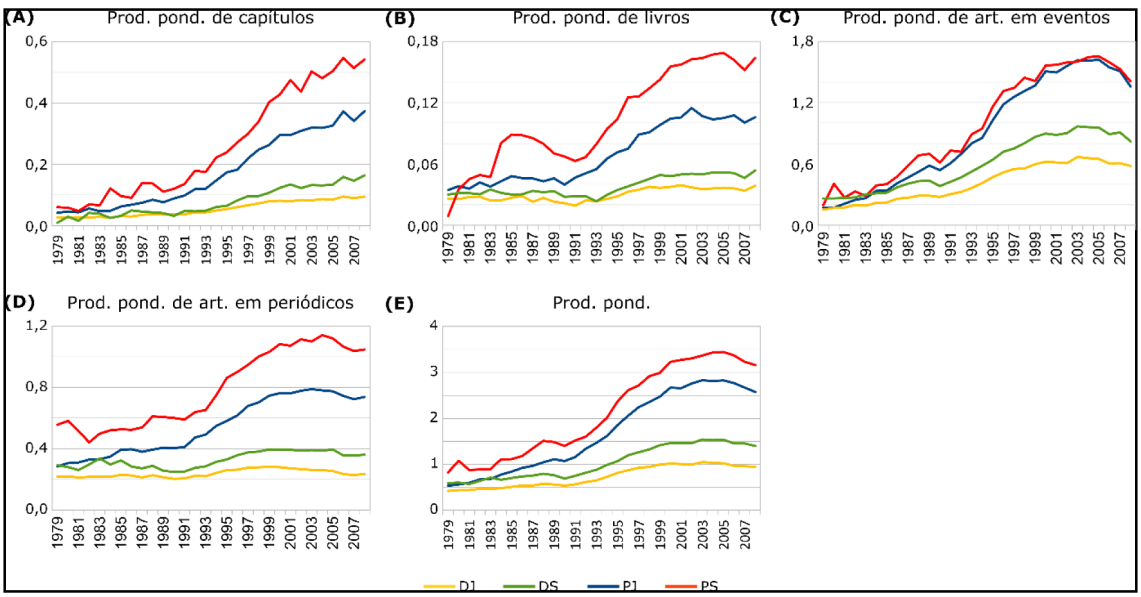
Esta seção aborda o número de publicações de artigos em eventos, periódicos, livros e capítulos de livros por tipo de acadêmico. Uma comparação é feita no sentido de verificar se um determinado tipo de acadêmico possui maior número de publicações em cotejo com os outros tipos. Para auxiliar essa verificação, a figura 6 apresenta a evolução da produtividade ponderada de acadêmicos, no período de 1979 até 2008, concernente a capítulos de livros (A), livros (B), artigos publicados em eventos (C) e em periódicos (D), bem como a produtividade ponderada total (considerando os quatro tipos de publicação, em E).

QP6: Há diferenças na quantidade e no tipo de publicações, considerando os papéis (professor ou descendente) exercidos por acadêmicos?

Como pode ser observado na **Figura 6**, para os quatro tipos de publicações, analisadas isoladamente (gráficos A-D) e em sua totalidade (gráfico E), a produtividade ponderada de PS é maior. E, como esperado (pelo tempo de carreira), a produtividade de PJ é maior que a de descendentes DJ e DS. Em outras palavras, isso significa que professores seniores possuem maior proporção de quantidade de publicações de artigos, capítulos de livros e livros, em comparação com as outras categorias e acadêmicos.

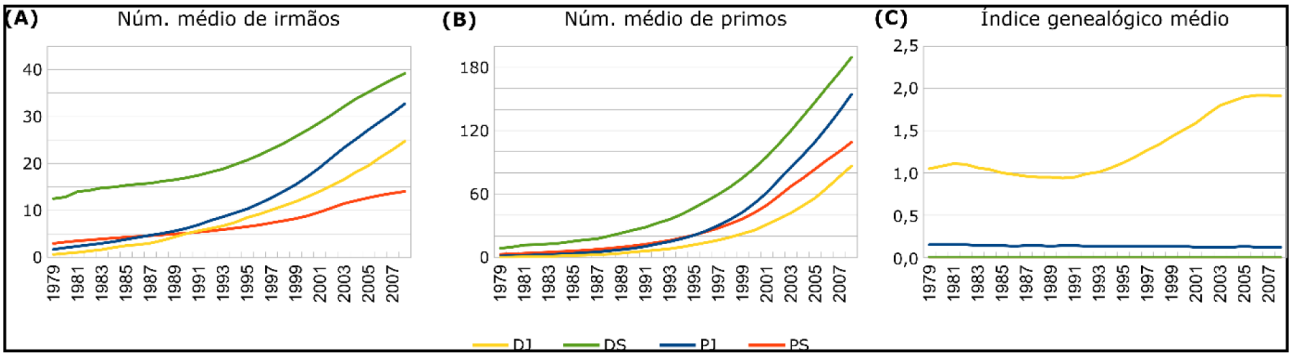
No contexto dos descendentes, cabe atentar-se para o fato de que a produtividade ponderada de DS é maior que a de DJ, isto é, descendentes orientados por professores mais experientes são mais produtivos em relação àqueles orientados por professores menos experientes. Na prática, esses são indícios de que alunos orientados por professores seniores publicam maior quantidade de artigos, capítulos de livros e livros, no cotejo com alunos orientados por professores juniores.

Figura 6 – Produtividade ponderada de acadêmicos total e por tipo de publicação no período de 1979 até 2008



Fonte: Elaborado pelos autores (2020).

Figura 7 – Número médio de irmãos, primos e índice genealógico de acadêmicos no período de 1979 até 2008



Fonte: Elaborado pelos autores (2020).

Analisar a rede desses acadêmicos pode elucidar uma das razões dessas diferenças de produtividade, dado que as publicações são realizadas, também, em coautoria. Nesse sentido, apresenta-se, na figura 7, o número médio de irmãos, primos e índice genealógico no período de 1979 até 2008, para os quatro tipos de acadêmicos.

Como esperado, o índice genealógico dos acadêmicos PS é maior que o dos outros três tipos de acadêmicos, e o de DJ e DS são semelhantes. Esse resultado é esperado, dado o tempo de carreira que esses acadêmicos possuem. No caso de professores menos experientes, é provável ainda não ter havido tempo hábil para ter outras gerações de descendentes (netos acadêmicos, por exemplo).

Com relação ao número médio de irmãos, nota-se que descendentes seniores possuem os maiores valores (ao longo de toda a série) e descendentes juniores os menores valores (a partir de 1991). Esse é um indício preliminar de que a maior rede dos acadêmicos DS pode influenciar a produção de publicações científicas. Esse fato também é verificado para o número médio de primos acadêmicos.

CONCLUSÕES

Este trabalho desenvolve um modelo populacional de crescimento para analisar grafos de genealogia acadêmica sob a ótica da ocorrência anual de orientações de mestrado e doutorado. Nesse modelo, os acadêmicos mudam de categoria de acordo com os seus respectivos históricos de orientação. Esse modelo então é aplicado aos dados de genealogia acadêmica do Brasil.

Os resultados evidenciam que uma fração cada vez menor de acadêmicos descendentes tem sido convertida em professores orientadores. Em termos bibliométricos, nota-se que professores seniores apresentam uma produtividade ponderada maior do que as demais categorias de acadêmicos. Da mesma forma, seus descendentes são mais produtivos do que os descendentes de professores juniores. Embora este trabalho não aprofunde os mecanismos subjacentes a essa observação, tal correlação de desempenho entre orientadores e orientandos poderia, em princípio, resultar de homofilia na formação das relações de orientação ou de transferência de habilidades inerentes à atividade de pesquisa dos orientadores para os seus orientados. Em trabalhos futuros, pretende-se: (a) empregar o método apresentado neste trabalho em outras bases de genealogia acadêmica; (b) analisar outras medidas de bibliometria, tais como citações; e (c) realizar análise dos dados segundo áreas do conhecimento.

REFERÊNCIAS

ALBERT, R.; JEONG, H.; BARABÁSI, A. L. Diameter of the world wide web. *Nature,* [*s. l.*], v. 401, p. 130-131, 1999. DOI: 10.1038/43601

BACAËR, N. *A short history of mathematical population dynamics*. London: Springer Science & Business Media, 2011. 160 p. ISBN: 978-0-85729-115-8. DOI: 10.1007/978-0-85729-115-8

CENTRO DE GESTÃO E ESTUDOS ESTRATÉGICOS (CGEE). *Mestres e doutores 2015:* estudos da demografia da base técnico-científica brasileira. Brasília/DF, 348 p., 2016. Disponível em: <https://www.cgee.org.br/web/rhcti/mestres-e-doutores-2015>. Acesso em: 11 mar. 2021.

DAMACENO, R. J. P. *et al*. The Brazilian academic genealogy: evidence of advisor–advisee relationships through quantitative analysis. *Scientometrics*, [*s. l.*], v. 119, n. 1, p. 303-333, 2019. DOI: 10.1007/s11192-019-03023-0

PRICE, D. J. de S. *Little science, big science*. New York: Columbia University Press, 1963. 119 p. ISBN: 9780231885751

ERDÖS, P.; RÉNYI, A. On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, v. 12, n. 1-2, p. 261-267, 1961. DOI: https://doi.org/10.1007/BF02066689

JACKSON, M. O.; ROGERS, B. W. Meeting strangers and friends of friends: how random are social networks?. *American Economic Review*, [*s. l.*], v. 97, n. 3, p. 890-915, 2007. DOI: 10.1257/aer.97.3.890

MALTHUS, T. R.; WINCH, D.; JAMES, P. *Malthus:* an essay on the principle of population. [*s. l.*]: Cambridge University Press, 1992. 430 p. ISBN-13 : 978-0521429726

ROSSI, L. *et al*. Topological metrics in academic genealogy graphs. *Journal of Informetrics*, [*s. l.*], v. 12, n. 4, p. 1042-1058, 2018. DOI: https://doi.org/10.1016/j.joi.2018.08.004

ROSSI, L; FREIRE, I. L.; MENA-CHALCO, J. P. Genealogical index: a metric to analyze advisor–advisee relationships. *Journal of Informetrics*, [*s. l.*], v. 11, n. 2, p. 564-582, 2017. DOI: https://doi.org/10.1016/j.joi.2017.04.001

WU, Y.; VENKATRAMANAN, S.; CHIU, D. A population model for academia: case study of the computer science community using DBLP bibliography 1960-2016. *IEEE Transactions on Emerging Topics in Computing,* [*s. l.*], v. 9, n. 1, p. 258-268, 2018. DOI: 10.1109/tetc.2018.2855156

Fusão de dados para análise de imagens registradas por satélites: proposta de modelo de metadados

Isaque Katahira
Doutorando em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp) - Palmital, SP - Brasil. Mestre em Bioinformática pela Universidade Tecnológica Federal do Paraná (UTFPR) - Brasil. Professor da Faculdade de Tecnologia (FATEC) - Pompeia, SP - Brasil.
http://lattes.cnpq.br/9340096523226667
https://orcid.org/0000-0001-5800-9890
E-mail: isaque.katahira@fatec.sp.gov.br

Danilo Camargo Dias
Mestrando em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp) - Marília, SP - Brasil. Especialização em Formação Pedagógica (Licenciatura) pela Centro Estadual de Educação Tecnológica Paula Souza (CEETEPS) - Brasil. Especialização em Engenharia de Sistemas pela Escola Superior Aberta do Brasil (ESAB) - Brasil. Professor da Escola Técnica Estadual (ETEC) - Itapeva, SP - Brasil.
http://lattes.cnpq.br/4387186879667385
https://orcid.org/0000-0002-9838-8861
E-mail: danilo.dias@etec.sp.gov.br

Danilo Dolci
Mestrando em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp) - Marília, SP - Brasil. Especialização em Pós-graduação em Redes de Computadores pela Faculdade Estácio de Sá de Ourinhos (FAESO) - Brasil. Especialização em Ciências da Computação pela Universidade Estadual de Londrina (UEL) – PR - Brasil. Professor e coordenador do Curso de Análise e Desenvolvimento de Sistemas da Faculdade de Tecnologia de Garça (FATEC) – Garça, SP - Brasil.
http://lattes.cnpq.br/1419975657681241
https://orcid.org/0000-0003-0415-6240
E-mail: danilo.dolci@fatec.sp.gov.br

Mariângela Spotti Lopes Fujita
Livre-docência pela Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP) - Brasil. Pós-Doutorado pela Universidad de Murcia (UM) - Espanha. Doutora em Ciências da Comunicação pela Universidade de São Paulo (USP) - São Paulo, SP - Brasil. Membro permanente do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp) - Marília, SP - Brasil.
http://lattes.cnpq.br/6530346906709462
http://orcid.org/0000-0002-8239-7114
E-mail: fujita@marilia.unesp.br

Leonardo Castro Botega
Pós-Doutorado pelo Instituto de Ciências Matemáticas e de Computação (ICMC-USP) - Brasil. Doutor em Ciência da Computação pela Universidade Federal de São Carlos (UFSCAR) – SP - Brasil. Professor da Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Brasil.
http://lattes.cnpq.br/6027755717265622
https://orcid.org/0000-0003-1495-5935
E-mail: leonardo.botega@unesp.br

Isidoro Gil Leiva
Doutor em Técnicas y Métodos Actuales en Información pela Universidad de Murcia (UM) - Espanha. Professor da Universidad de Murcia (UM) - Campus de Espinardo - Murcia - Espanha
http://lattes.cnpq.br/4334024093986852
https://orcid.org/0000-0002-7175-3099
E-mail: isgil@um.es

Submetido em: 01/12/2020. Aprovado em: 09/03/2021. Publicado em: 28/07/2021.

RESUMO

Diversas tecnologias surgiram nas últimas décadas, que propiciaram um crescimento considerável no volume e na complexidade dos repositórios digitais de imagem. Diante desse desenvolvimento, também se faz necessária a criação de novas estratégias de tratamento da informação, sobretudo aquelas ligadas à organização e recuperação. Historicamente, o uso de metadados é uma técnica utilizada para identificação de arquivos de imagem. Nesse sentido, a fusão de dados pode ser utilizada como um método para conectar parâmetros em conjuntos de dados heterogêneos para indexação em repositórios de imagem com vistas a estabelecer conexões. Sendo assim, o objetivo deste trabalho é demonstrar as potencialidades da fusão de dados para a organização de imagens registradas por diferentes satélites, especificamente, desenvolver um modelo de metadados para a representação da informação na fusão de dados por sistemas de IA. Esta pesquisa é uma revisão bibliográfica exploratória, utilizada para demonstrar as potencialidades da fusão de dados a partir de imagens registradas por diferentes satélites, oriundas de dois sistemas de monitoramento da NASA - *National Snow & Ice Data Center* e o *GES DISC Data Archive* -, com posterior desenvolvimento de proposta de modelo de metadados para auxiliar a fusão de dados através da observação de parâmetros comuns e da exploração da estrutura dos metadados. Ao concluir este trabalho, pôde-se comprovar que a fusão de dados de imagem se destaca como uma estratégia eficiente de representação e recuperação da informação e colabora com a ampliação da compreensão das alterações climáticas no planeta.

Palavras-chave: Repositórios digitais de imagem. Metadados de imagens. Ciência de dados.

Fusion of data for analysis of images recorded by satellites: proposal of metadata model

ABSTRACT

Several technologies have emerged in recent decades that have provided considerable growth in the volume and complexity of digital image repositories. In view of this development, it is also necessary to create new information treatment strategies, especially those related to organization and recovery. Historically, the use of metadata is a technique used to identify image files. In this sense, data fusion can be used as a method to connect parameters in heterogeneous data sets for indexing in image repositories, in order to establish connections. In this sense, the objective of this work is to demonstrate the potential of data fusion for the organization of images recorded by different satellites, specifically, to develop a metadata model for the representation of information in data fusion by AI systems. This research is an exploratory bibliographic review used to demonstrate the potential of data fusion from images recorded by different satellites, coming from two NASA monitoring systems: National Snow & Ice Data Center and the GES DISC Data Archive, with further development proposal of a metadata model to assist data fusion, by observing common parameters and exploring the metadata structure. At the end of this work, it was possible to prove that the fusion of image data stands out as an efficient strategy for the representation and recovery of information and collaborates to expand the understanding of climate change on the planet.

Keywords*: Digital image repository. Image metadata. Data science.*

Fusión de datos para el análisis de imágenes grabadas por satélites: propuesta de modelo de metadatos

Resumen

Han surgido varias tecnologías en las últimas décadas que han proporcionado un crecimiento considerable en el volumen y la complejidad de los repositorios de imágenes digitales. En vista de este desarrollo, también es necesario crear nuevas estrategias de tratamiento de la información, especialmente las relacionadas con la organización y la recuperación. Históricamente, el uso de metadatos es una técnica utilizada para identificar archivos de imagen. En este sentido, la fusión de datos puede usarse como un método para conectar parámetros en conjuntos de datos heterogéneos para indexar en repositorios de imágenes, a fin de establecer conexiones. Mientras tanto, el objetivo de este trabajo es demostrar el potencial de fusión de datos para la organización de imágenes grabadas por diferentes satélites, específicamente, para desarrollar un modelo de metadatos para la representación de información en fusión de datos por sistemas de IA. Esta investigación utiliza la revisión bibliográfica exploratoria para demostrar el potencial de fusión de datos de imágenes grabadas por diferentes satélites, provenientes de dos sistemas de monitoreo de la NASA: el Centro Nacional de Datos de Nieve y Hielo y el Archivo de Datos GES DISC, con un mayor desarrollo propuesta de un modelo de metadatos para ayudar a la fusión de datos, observando parámetros comunes y explorando la estructura de metadatos. Al final de este trabajo, fue posible demostrar que la fusión de datos de imágenes se destaca como una estrategia eficiente para la representación y recuperación de información y colabora para expandir la comprensión del cambio climático en el planeta.

***Palabras clave**: Repositorios de imágenes digitales. Metadatos de imagen. Ciencia de datos.*

INTRODUÇÃO

Com o crescimento dos repositórios digitais de imagens, as técnicas de fusão de dados têm conquistado uma importância cada vez maior. A intensidade e a heterogeneidade com que arquivos, especialmente os de imagens, são produzidos, traz novos e grandes desafios, pois à medida que os dados gerados se multiplicam, há também a necessidade crescente de criar estratégias para sua organização e recuperação. Uma forma de aprimorar a descrição e a identificação de documentos de imagens é a utilização de metadados.

O indexador deve considerar tanto a rapidez quanto a precisão proporcionada pelos metadados catalogados. No entanto, observando diferentes fontes, percebe-se que nem sempre os metadados são convergentes, fato que pode trazer dúvidas aos usuários ou explicitar lacunas de observação. Nesse sentido, a fusão de dados configura-se como uma estratégia importante para extração de parâmetros mais representativos (HALL; JORDAN, 2010; BOTEGA, 2016).

Para Juan e Tauler (2019), a fusão de dados implica com frequência na concatenação de conjuntos de dados que apresentam uma enorme diversidade em termos de informação, tamanho e comportamento. As informações conectadas refletem a variação atribuída por componentes, eventos ou fontes, agregando dados representativos que podem ser analisados simultaneamente.

É notório que as últimas décadas presenciaram o surgimento de novas aplicações para a tecnologia da informação, sobretudo no que tange à organização de dados dos repositórios de imagens digitais. Somado a isso, o compartilhamento de documentos aumentou, principalmente, quando se trata de fotografias. Contudo, segundo Oliveira e Vital (2015), não há, na literatura científica da área da Ciência da Informação, um acordo sobre como as imagens devem ser tratadas, pois diversos fatores implicam nessa análise, incluindo o tipo de imagem (pinturas, fotografias etc.) e o suporte no qual ela é divulgada.

Nesse contexto, a hipótese que orientará a investigação realizada é que a execução do registro de um fenômeno a partir de diversos critérios colabora para a sua compreensão mais global. Assim, a fusão de dados se mostraria relevante à obtenção de um resultado superior de recuperação, na medida em que viabilizaria a comparação, a extração e a correlação de dados sob diferentes perspectivas.

A fim de confirmar a validade da hipótese formulada, utiliza-se a Ciência de Dados para reunir e comparar informações sobre eventos ambientais, bem como identificar os registros mais significativos para sua recuperação. A proposição desta pesquisa é evidenciar as potencialidades da fusão de dados para a organização de imagens registradas por diferentes satélites e, assim, desenvolver um modelo de metadados para a representação da informação na fusão de dados por sistemas de inteligência artificial (IA). As imagens selecionadas são resultado do monitoramento do dióxido de carbono livre na troposfera, da temperatura dos oceanos e da temperatura atmosférica.

A relevância da proposta se confirma pela inegável utilidade das imagens de satélite para monitoramento de grandes áreas, permitindo a compreensão e a avaliação de alterações ambientais (tanto positivas quanto negativas), ao correlacionar, por exemplo, os índices de emissão de CO_2 (dióxido de carbono) à temperatura média dos oceanos e à temperatura atmosférica. Ademais, a atualidade temática confirma a validade da pesquisa, já que iniciativas de aprimorar o uso da ciência em benefício do meio ambiente se mostram urgentes.

Nesse ínterim, a fusão de dados revela-se fundamental ao potencializar as informações obtidas pela ciência, permitindo avaliações capazes de orientar políticas baseadas em evidências, a fim de contribuir para a proposição de ações efetivas que auxiliem na manutenção da vida humana (SRIVASTAVA *et al.*, 2019; VENKATESH *et al.*, 2019).

Para alcançar os objetivos pretendidos, inicialmente, realiza-se revisão bibliográfica exploratória sobre padrões de metadados e fusão de dados de imagens. Na sequência, consulta-se dois sistemas de monitoramento da NASA, o *National Snow & Ice Data Center* e o *GES DISC Data Archive*, especificamente com relação aos metadados catalogados sobre as imagens de interesse, para, finalmente, propor a fusão dos dados encontrados com vistas a melhor compreender os efeitos dos altos índices de emissão de CO_2 para a qualidade de vida no planeta.

FUNDAMENTAÇÃO TEÓRICA

A priori, a revisão bibliográfica exploratória subsidiou a pesquisa por permitir o desenvolvimento de hipóteses, o levantamento da fortuna crítica disponível e a compreensão organizada dos conceitos (MARCONI; LAKATOS, 2017). De modo preliminar, para destacar a importância da fusão de dados como estratégia que congrega informações relevantes, é oportuno mencionar a criação de metadados.

Segundo Alves (2012), os metadados são dados que possibilitam a recuperação, descrição, avaliação, interpretação e manipulação de documentos eletrônicos. Nesse sentido, os metadados são considerados essenciais em diversas fases do ciclo informacional. Para Simionato (2012), não há como recuperar ou preservar um recurso informacional sem que exista um registro com base em metadados, pois são eles que representam documentalmente os arquivos selecionados.

No âmbito das produções de imagens, o registro de metadados para identificação e recuperação de informações é ainda mais desafiador, tendo em vista a necessidade de registrar, por meio de palavras, dados obtidos por intermédio da observação da linguagem não verbal (YAMANE; CASTRO, 2018).

Há diferentes formas de representação ou reprodução de uma imagem: pinturas, ilustrações, animações, ícones, fotografias entre outras.

Sobre a fotografia, pode-se defini-la como um gênero textual que apresenta certa complexidade para o processo de indexação, uma vez que é o resultado da captação e fixação de uma determinada ação, em determinado momento e lugar, por determinado agente. Diante dessa captação, muitíssimo peculiar, o indexador encontra a difícil tarefa de registrar verbalmente o que fora registrado com todas as nuances da linguagem não verbal. Diante desse desafio, fica evidente a interferência da subjetividade, pois, se a captação foi feita em determinado tempo-espaço, também o contexto da leitura feita pelo indexador é relevante e pode interferir nos termos atribuídos durante a indexação.

Dessa forma, o uso de metadados e descritores em imagens fotográficas mostra-se fundamental, visto que existem informações significativas nesse tipo de conteúdo que podem ser perdidas, caso os processos de catalogação e indexação não sejam realizados conforme os critérios validados.

> A catalogação e a indexação são mecanismos que condensam a informação, a fim de sua representação, para a sua recuperação. A partir disso, a catalogação e a indexação de fotografia em ambientes digitais se fazem importantes, não podendo ser elaboradas de qualquer maneira (FELIPE; FELIPE, 2016, p. 5).

Ainda de acordo com Felipe e Felipe (2016), a propósito da indexação de imagens, é necessário manter em perspectiva que esse tipo de registro é carregado de subjetividade e, por isso, sua descrição por meio de metadados é essencial à recuperação de seu conteúdo.

Na esfera científica, as fotografias obtidas a partir do sensoriamento remoto se materializam como valiosas fontes de informação para pesquisas relacionadas ao clima do planeta (ZANOTTA; FERREIRA; ZORTEA, 2019). Recortando intervalos espaço-temporais, as fotografias permitem tanto o acompanhamento sincrônico quanto diacrônico das questões ambientais, fato que permite um olhar contextualizado sobre as principais mudanças naturais e os fenômenos provocados pela interação humana.

Segundo Zanotta, Ferreira e Zortea (2019), a crescente facilidade para se conseguir imagens da superfície terrestre em alta resolução, graças ao uso de satélites, possibilitou uma aproximação inovadora entre a tecnologia e a sociedade. As imagens fixadas trazem um imenso conteúdo informacional que, coletado e processado continuamente em diversas frentes de pesquisa científica, pode contribuir de maneira decisiva para a compreensão da vida no planeta.

Chino, Romani e Traina (2010) afirmam que, como os dados são gerados por diversos tipos de satélites, com imagens registradas em resolução e periodicidades diferentes, é de suma importância investir na integração das informações obtidas. Para que essa integração se concretize exponencialmente, a fusão de dados – conceituada como um conjunto de processos contínuos de refinamento e avaliações que lidam com múltiplos dados e informações, de fontes únicas ou diversas – é uma estratégia vital (HALL; JORDAN, 2010).

Segundo Botega (2016), a fusão de dados ou informações corresponde à rotina de transformação de dados ou informações para produzir estimativas e predições de estados de entidades, visando a maximizar o valor da informação e estimular a consciência situacional de analistas sobre um ambiente de interesse.

Sobre a questão, a título de exemplo, o trabalho de Srivastava *et al*. (2019) busca detectar áreas com alta probabilidade de incêndio florestal em uma região da Índia, com base na diferença da *Normalized Burn Ratio* (NBR), entre as condições pré e pós-incêndio. O estudo compara *datasets* gerados pelo satélite LANDSAT TM 5[1] por dois modelos de captura, o *Geographical Information Systems* (GIS)[2] e o *Earth Observation* (EO)[3]. O resultado é uma avaliação multicritério, incorporando atributos como fontes antropogênicas e naturais, de modo a fundir os modelos e realizar a predição de zonas com alta probabilidade de incêndio.

[1] Disponível em: http://landsat.gsfc.nasa.gov/. Acesso em: 17 jun. 2020.
[2] Disponível em: https://grass.osgeo.org/. Acesso em: 17 jun. 2020.
[3] Disponível em: https://earthobservatory.nasa.gov/. Acesso em: 17 jul. 2020.

A partir do exposto, são de suma relevância estudos que abordem a fusão de dados em relação a todas as áreas do conhecimento científico, sobretudo no tratamento de imagens fotográficas. A coleta e a organização de informações oriundas de diversas fontes configuram-se como estratégia eficiente em busca da qualidade e da assertividade na tomada de decisões.

METODOLOGIA

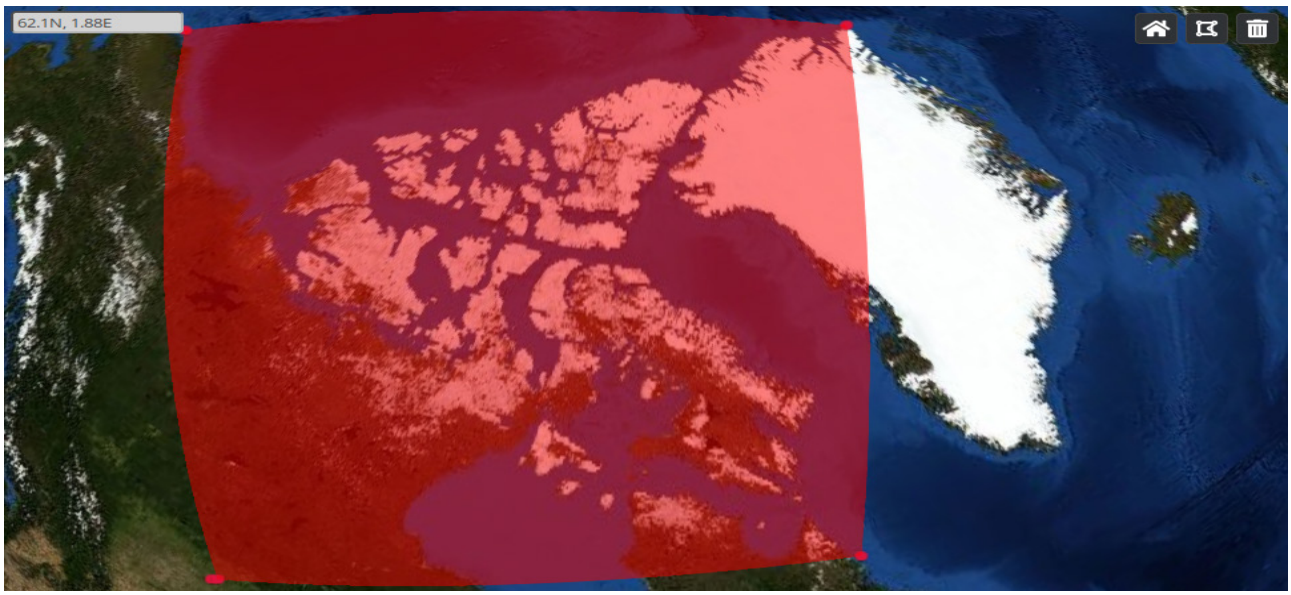
Com o intuito de confirmar a fusão de dados como ferramenta capaz de reunir, selecionar e evidenciar dados relevantes, é realizada a consulta e seleção de imagens e metadados de dois sistemas de monitoramento da *National Aeronautics and Space Administration* (NASA)[4], o *National Snow & Ice Data Center* (NSIDC)[5] e o *GES DISC Data Archive*[6], especificamente com relação à temática de interesse.

A partir dos registros selecionados, evidencia-se as potencialidades da fusão dos dados encontrados, a fim de melhor compreender como os altos índices de emissão de CO2 reverberam a qualidade de vida na Terra.

As ferramentas *National Snow & Ice Data Center* e *GES DISC Data Archive* foram acessadas pela área de dados abertos da NASA para extrair metadados de imagens obtidas por satélites que monitoram o dióxido de carbono livre na troposfera, a temperatura dos oceanos e a temperatura atmosférica**.**

O NSIDC é um centro de pesquisas da NASA que gerencia e distribui dados científicos sobre a criosfera, isto é, regiões da superfície terrestre cobertas permanentemente por gelo e neve. Aos usuários cadastrados, o sistema permite a solicitação de informações de uma área que deve ser delimitada pelo requerente. Diversas informações são geradas e compactadas num arquivo que fica disponível na área do usuário, conforme figura 1.

Figura 1 – A área destacada em vermelho é a localização da qual foram extraídos os metadados



Fonte: National Snow & Ice Data Center, 2019.

[4] Disponível em: https://www.nasa.gov/. Acesso em: 17 jul. 2020.
[5] Disponível em: https://nsidc.org/. Acesso em: 17 jul. 2020.
[6] Disponível em: https://disc.gsfc.nasa.gov/. Acesso em: 17 jul. 2020.

Ao se consultar uma imagem fotográfica disponibilizada pelo NSIDC em seu ambiente digital, pode-se ver que são utilizados metadados para as descrições. Os campos que compõem a catalogação são: identificador do *datacenter*, data de registro, data da última alteração; nome do arquivo compartilhado; tamanho (MB); data e hora da captura; espaço de tempo inicial da captura; espaço de tempo final da captura; coordenadas da área observada (área retangular - norte, sul, leste e oeste); nome da plataforma; instrumento de sensoriamento utilizado; nome da campanha da NASA e identificador da aeronave utilizada.

A partir da figura 1, é possível obter treze registros de metadados, explicitados no quadro 1.

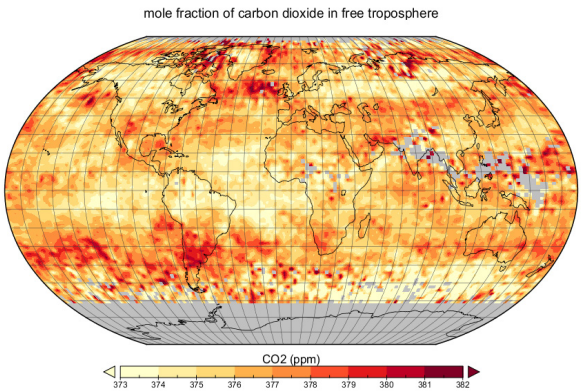
Quadro 1 − Derretimento das calotas polares

National Snow & Ice Data Center	
Identificador do datacenter	NSIDC
Data de registro	2017-08-07T15:50:19Z
Data da última alteração	2017-08-10T15:23:31Z
Nome do arquivo compartilhado	ILATM1B_20180419_181100.ATM6AT6.h5
Tamanho (MB)	5.62564
Data e hora da captura	2017-12-18 03:18:19.000
Espaço de tempo inicial da captura	2017-04-19T18:11:00Z
Espaço de tempo final da captura	2017-04-20T08:25:00Z
Coordenadas da área observada (Área retangular – Norte, Sul, Leste e Oeste)	76.54604 76.53789 76.53673 76.54488
Nome da plataforma	P-3B
Instrumento de sensoriamento utilizado	ATM
Nome da campanha da NASA	2017_GR_NASA
Identificador da aeronave utilizada	N426NA

Fonte: Elaborado pelo autor com base nos dados obtidos na ferramenta *National Snow & Ice Data Center, 2019.*

O *GES DISC Data Archive* é o sistema que fornece dados, informações e serviços de ciências da Terra para pesquisadores, cientistas de dados, usuários de aplicativos e estudantes, incluindo: composição atmosférica, ciclos de água e energia, e variabilidade climática. É possível observar que também são arquivados conjuntos de dados aplicáveis ao Ciclo de Carbono e ao Ecossistema. Assim, o *GES DISC* disponibiliza diversos *datasets* sobre a situação climática do planeta. O sistema gera e renderiza imagens e metadados sobre regiões em diversos padrões concernentes à emissão de CO_2, à temperatura dos oceanos e à temperatura atmosférica. Quanto à emissão de CO_2, selecionou-se a seguinte imagem:

Figura 2 − Dióxido de carbono livre na troposfera



Fonte:https://docserver.gesdisc.eosdis.nasa.gov/public/project/Images/AIRS3C28_005.png, 2019.

Sobre a figura 2, os treze metadados relacionados na catalogação são: data de registro; data da última alteração; nome da coleção a que pertence; identificador da coleção; tamanho (MB); período (diurno / noturno); data e hora da captura; espaço de tempo inicial da captura; espaço de tempo final da captura; coordenadas da área observada (área retangular - norte, sul, leste e oeste); parâmetros medidos; parâmetro de qualidade automatizada e URL de acesso online ao recurso. Nesta pesquisa, optou-se pelo padrão Nativo, que permitiu a identificação dos metadados, conforme quadro 2.

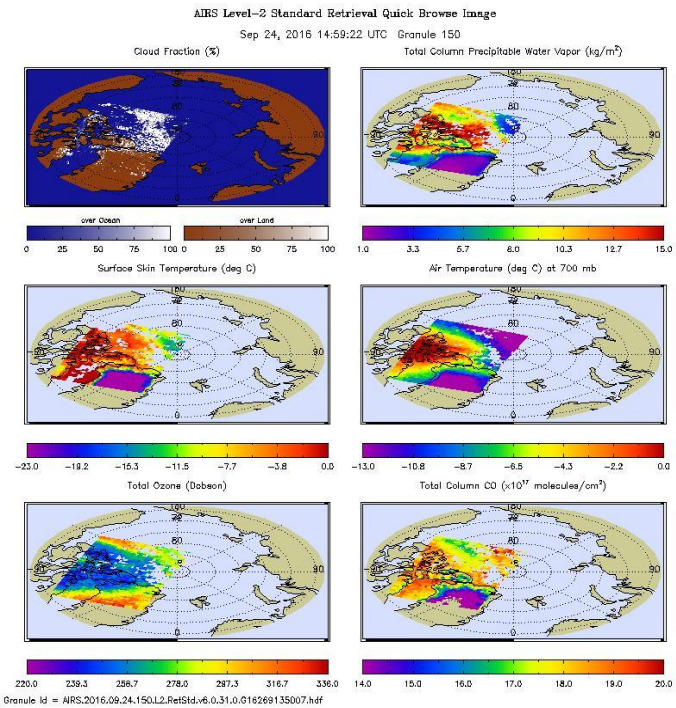
No que tange à temperatura dos oceanos, selecionou-se, a título de exemplo, a figura 3, com dezessete metadados descritos.

Quadro 2 – Emissão de CO_2

GES DISC Data Archive	
Data de registro	2017-08-07T15:50:19Z
Data da última alteração	2017-08-09T15:50:19Z
Nome da coleção a que pertence	AIRS3C28
Identificador da coleção	5
Tamanho (MB)	0.503100395202637
Período (Diurno/Noturno)	Day
Data e hora da captura	2017-03-09T22:51:31.000Z
Espaço de tempo inicial da captura	2017-03-08T00:00:00.000000Z
Espaço de tempo final da captura	2017-03-06T23:59:59.999999Z
Coordenadas da área observada (Área retangular - Norte, Sul, Leste e Oeste)	-180.0 90.0 180.0 -90.0
Parâmetros medidos	Carbon Dioxide Column Density
Parâmetro de qualidade automatizada	Passed
URL de acesso online ao recurso	http://acdisc.gesdisc.eosdis.nasa.gov/data//Aqua_AIRS_Level3/AIRS3C28.005/2017/AIRS.2017.02.14.L3.CO2Std_IR008.v5.9.14.0.IRonly.X17068145128.hdf

Fonte: Elaborado pelo autor com base nos dados obtidos na ferramenta GES DISC Data Archive, 2019.

Figura 3 – Temperatura dos oceanos



Fonte:https:///airsl2.gesdisc.nasa.gov/data/Aqua_AIRS_Level2/AIRX2RET.006/2016/268/AIRS.2016.09.24.150.L2.RetStd.v6.0.31.0.G16269135007.hdf.jpg, 2019.

Sobre a figura 3, os metadados relacionados na catalogação são: data de registro; data da última alteração; nome da coleção a que pertence; indicador da coleção; tamanho (MB); tipo de camada atmosférica observada; período (diurno/noturno); data e hora da captura; espaço de tempo inicial da captura; espaço de tempo final da captura; domínio espacial vertical (máximo e mínimo); coordenadas da área observada (área retangular - norte, sul, leste e oeste); URL de acesso online ao recurso, descrição; tipo de dado (aberto ou não); *mime type* e navegação em imagens associadas (URL, tamanho e descrição), como registrado no quadro 3.

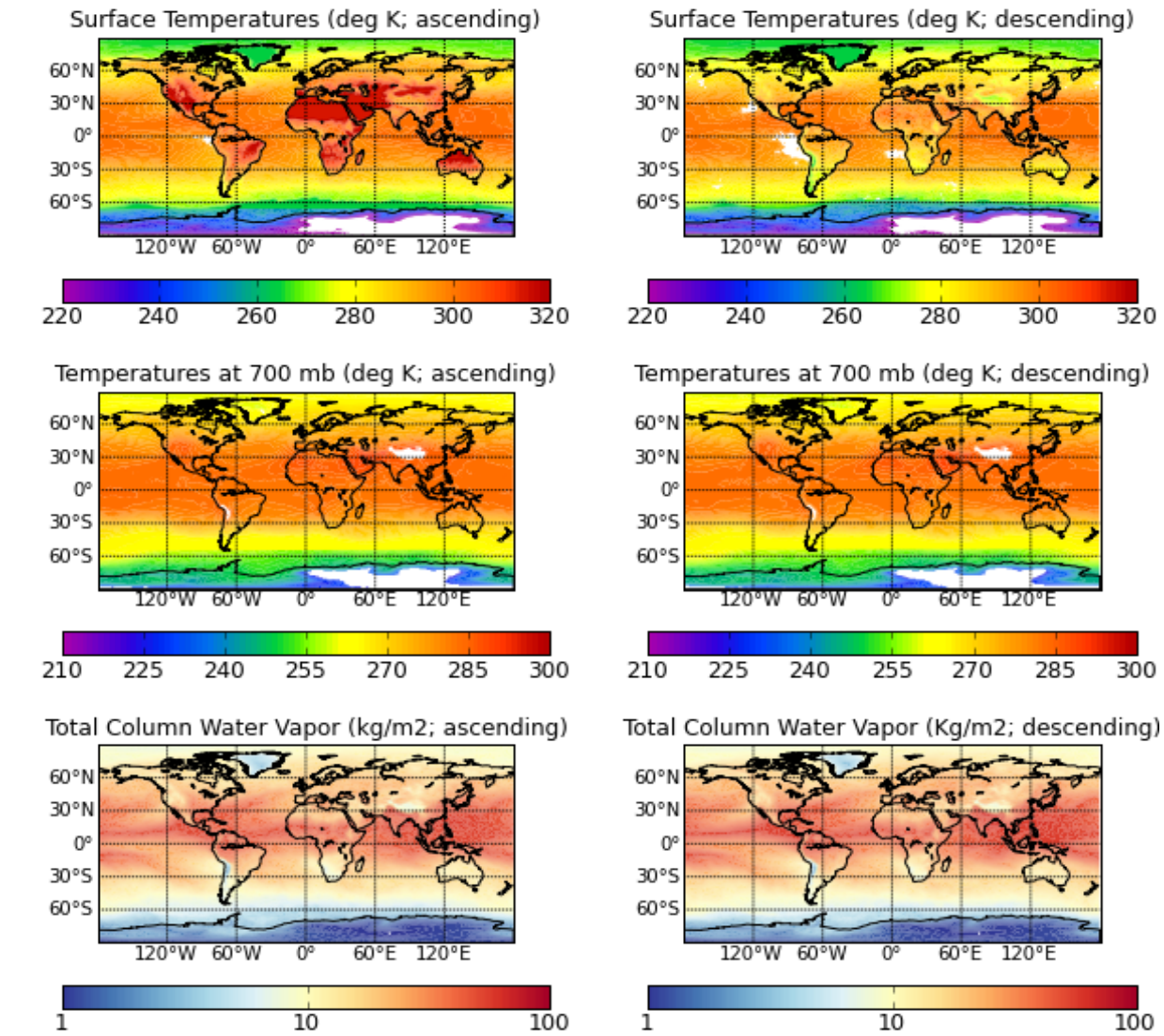
A última imagem selecionada para apresentação de metadados representativos faz referência à temperatura atmosférica.

Quadro 3 – Temperatura dos oceanos

GES DISC Data Archive	
Data de registro	2017-08-07T15:50:19Z
Data da última alteração	2017-08-08T15:54:19Z
Nome da coleção a que pertence	AIRX2RET
Identificador da coleção	6
Tamanho (MB)	3.583529472351074
Tipo de camada atmosférica observada	Atmosphere Profile
Período (Diurno/Noturno)	BOTH
Data e hora da captura	2017-04-25T17:50:08.000Z
Espaço de tempo inicial da captura	2017-04-24T14:59:22.000000Z
Espaço de tempo final da captura	2017-04-25T14:59:22.000000Z
Domínio espacial vertical (máximo e mínimo)	0.1mb, or appr. 50km
Coordenadas da área observada (Área retangular – Norte, Sul, Leste e Oeste)	160.906860351562 90.0 -22.1724796295166 68.8115463256836
URL de acesso online ao recurso	https://airsl2.gesdisc.eosdis.nasa.gov/data/Aqua_AIRS_Level2/AIRX2RET.006/2016/268/AIRS.2016.09.24.150.L2.RetStd.v6.0.31.0.G16269135007.hdf
Descrição	The OPENDAP location for the granule
Tipo de dado (Aberto ou não)	GET DATA : OPENDAP DATA
Mime type	application/x-hdf
Navegação em imagens associadas (URL, tamanho e descrição)	https://airsl2.gesdisc.eosdis.nasa.gov/data/Aqua_AIRS_Level2/AIRX2RET.006/2016/268/AIRS.2016.09.24.150.L2.RetStd.v6.0.31.0.G16269135007.hdf.jpg 170797 Browse image for AIRS.2016.09.24.150.L2.RetStd.v6.0.31.0.G16269135007.hdf

Fonte: Elaborado pelo autor com base nos dados obtidos na ferramenta GES DISC Data Archive, 2019.

Figura 4 – Temperatura atmosférica



Fonte: https://docserver.gesdisc.eosdis.nasa.gov/public/project/Imag es/AIRH3SPM_006.png, 2019.

Para descrição da figura 4, são fixados quinze metadados: data de registro; data da última alteração; nome da coleção a que pertence; identificador da coleção; tamanho (MB); período (diurno / noturno); data e hora da captura; espaço de tempo inicial da captura; espaço de tempo final da captura; domínio espacial vertical (máximo e mínimo); coordenadas da área observada (área retangular - norte, sul, leste e oeste); URL de acesso online ao recurso; descrição; tipo de dado (aberto ou não) e *mime type*, como registrado no quadro 4:

Quadro 4 – Temperatura atmosférica

GES DISC Data Archive	
Data de registro	2017-08-07T15:50:19Z
Data da última alteração	2017-08-09T15:55:32Z
Nome da coleção a que pertence	AIRIBQAP_NRT
Identificador da coleção	5
Tamanho (MB)	2.480497360229492
Período (Diurno/ Noturno)	DAY
Data e hora da captura	2017-05-24T13:26:06.000Z
Espaço de tempo inicial da captura	2017-05-23T12:29:21.000000Z
Espaço de tempo final da captura	2017-05-24T10:35:20.999999Z
Domínio espacial vertical (máximo e mínimo)	TOA SFC
Coordenadas da área observada (Área retangular – Norte, Sul, Leste e Oeste)	-98.5035018920898 90.0 17.6566963195801 63.0190010070801
URL de acesso online ao recurso	https://discnrt1.gesdisc.eosdis.nasa.gov/data//Aqua_AIRS_NRT/AIRIBQAP_NRT.005/2019/236/AIRS.2019.08.24.125.L1B.AIRS_QaSub.v5.0.23.0.R19236092606.hdf
Descrição	The OPENDAP location for the granule
Tipo de dado (aberto ou não)	GET DATA: OPENDAP DATA
Mime type	application/x-hdf

Fonte: Elaborado pelo autor com base nos dados obtidos na ferramenta GES DISC Data Archive, 2019.

Assim, tendo destacado as imagens selecionadas para a amostra e exposto os respectivos metadados, apresenta-se a análise dos resultados e um modelo para a fusão de dados.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

Diante da seleção das imagens e da enumeração-descritiva dos respectivos metadados apresentados nos quadros 1 a 4, é possível constatar que a complementaridade das informações catalogadas permite, ao pesquisador, maior abrangência de análise do fenômeno de interesse. Sobre o derretimento das calotas polares, a identificação de diversas variáveis, possivelmente associadas, tais como, a concentração de CO2 na troposfera, a temperatura dos oceanos e a temperatura atmosférica, permite um olhar mais sistêmico sobre o problema.

Após a seleção e identificação das imagens da amostra, a organização da fusão de dados foi realizada em três etapas: i) verificação dos metadados que uma base e outra utilizam, ou seja, exploração da estrutura dos metadados, da quantidade, da descrição, do tipo entre outros; ii) verificação da seleção de termos comuns – quantidade e qualidade –, a fim de explicitar os parâmetros de fusão de dados selecionados; e iii) seleção das imagens após a realização da fusão de dados, com posterior tabulação para análise comparada.

A propósito da última etapa, cumpre destacar que a comparação se deu por meio da análise dos dados obtidos separadamente em relação aos dados reunidos.

Pela seleção e descrição das imagens, por intermédio dos metadados identificados, é possível gerar uma visão global da qualidade de vida na Terra. O quadro 5 exibe uma proposta de fusão de dados. Para tanto, foi selecionada uma área retangular do norte do globo até o extremo sul, de modo a fixar esse eixo.

Concernente à fusão de dados, destacando os parâmetros comuns e a verificação das estratégias que uma e outra utilizam de metadados, ou seja, exploração da estrutura dos metadados, da quantidade, da descrição, do tipo entre outros, os resultados, na primeira etapa, são apresentados conforme disposto no quadro 5.

Com base no quadro 5, é possível inserir os dados em um determinado algoritmo de IA (OSÓRIO; BITTENCOURT, 2000) que faça a leitura e a interpretação dos dados, em um determinado período, e posteriormente, estabeleça a correlação existente entre eles. Para tanto, uma técnica potencial é a de redes neurais, que são amplamente utilizadas em softwares de análise de imagens (GONÇALVES *et al*., 2008; MOREIRA *et al.*, 2002). Destaca-se o Modelo de *Hopfield* de redes neurais artificiais (RNA) para analisar as imagens obtidas por satélites, haja vista apresentar os conceitos de redes com realimentação e comportamento dinâmico (HAYKIN, 2007).

Outra proposta é a utilização do algoritmo de MAXVER, ou Máxima Verossimilhança, que, pela classificação supervisionada, "[...] considera a ponderação das distâncias entre as médias dos níveis digitais das classes e o pixel, utilizando parâmetros estatísticos, isto é, considerando a distribuição de probabilidade normal para cada classe [...]" (RIBEIRO; CENTENO, 2001, p. 1342).

Nesse sentido, o algoritmo especialista de IA analisa a poluição do ar e a temperatura dos oceanos, estabelecendo, dessa forma, uma correlação com a elevação da temperatura global. Tal procedimento acrescenta novas informações aos metadados das imagens e viabiliza uma série de análises automatizadas da situação do planeta embasando-se na fusão dos dados (que estão separados e armazenados nos metadados) das imagens.

Quadro 5 − Modelo para visão global da qualidade do planeta Terra a partir da fusão de parâmetros

National Snow & Ice Data Center e GES DISC Data Archive
Data de registro
Data da última alteração
Nome da coleção a que pertence
Identificador da coleção
Tamanho (MB)
Período (Diurno/Noturno)
Data e hora da captura
Espaço de tempo inicial da captura
Espaço de tempo final da captura
Coordenadas da área observada (Área retangular – Extremo Norte Global Fixxo, Extremo Sul Global Fixo, Leste Variável e Oeste Variável)
Quantidade de CO2 registrado
Camada atmosférica onde foi registrada a medição da temperatura dos oceanos
Registro da temperatura média atmosférica continental
Dimensão das calotas polares registradas na faixa capturada
National Snow & Ice Data Center e GES DISC Data Archive
Correlação entre as médias das temperaturas aferidas
Correlação da quantidade de CO2 e a temperatura média atmosférica
Correlação da dimensão das calotas polares e a temperatura média atmosférica
URL de acesso aos recursos imagéticos

Fonte: Elaborado pelo autor com base nos dados obtidos nas ferramentas *National Snow & Ice Data Center e a GES DISC Data Archive, 2019.*

CONSIDERAÇÕES FINAIS

Em meio à heterogeneidade, ao volume e à complexidade de dados disponibilizados diariamente, as estratégias para seleção de informações têm ganhado destaque. No cenário da produção de imagens, essa realidade não é diferente, visto que o volume de dados proporcionados pela comunidade em geral, embora impulsione avanços em todas as áreas do saber, também torna necessário o desenvolvimento de estratégias para sua seleção e utilização eficiente.

Diante das imagens e dos metadados apresentados com foco no monitoramento de grandes áreas, foi possível ampliar a compreensão e a avaliação de alterações climáticas (tanto positivas quanto negativas), ao correlacionar, por exemplo, os índices de emissão de CO_2 à temperatura dos oceanos e à temperatura atmosférica.

Considerando as potencialidades dos novos recursos disponibilizados pelas ferramentas computacionais, foi possível comprovar que o registro de um fenômeno feito com base em diversos critérios colabora com a sua compreensão mais global, como proposto pela utilização da fusão de dados. A propósito das limitações da pesquisa com imagens, tem-se as restrições técnicas dos sensores, as dificuldades de descrição verbal e possíveis ruídos na catalogação.

Assim, a fusão de dados da área Espacial nas bases de dados do repositório da NASA mostrou-se relevante para a obtenção de um resultado superior de recuperação, pois evidenciou a complementaridade dos dados a partir das correlações estabelecidas. Logo, foi possível destacar os registros mais significativos para sua recuperação.

Por fim, convém sublinhar que a fusão de dados de imagens e os novos métodos de seleção emergem no cenário atual como estratégias eficientes de representação e recuperação da informação. Para trabalhos futuros, o método poderia ser testado em um *corpus* maior de imagens, bem como em outras áreas do conhecimento, de modo a ampliar a validação do comparativo, incluindo análise por meio de RNAs.

REFERÊNCIAS

ALVES, C. D. Metadados para a Recuperação de Imagens na WEB: utilizando o software ADOBE BRIDGE. *PontodeAcesso,* Salvador, v. 6, n. 1, p. 32-48, 2012.

DOI: http://dx.doi.org/10.9771/1981-6766rpa.v6i1.5131. Disponível em: https://periodicos.ufba.br/index.php/revistaici/article/view/5131. Acesso em: mar. 2021.

BOTEGA, L.C. *Modelo de fusão dirigido por humanos e ciente de qualidade de informação.* Orientadora: Regina Borges de Araújo. 2016. Tese (Doutorado em Ciência da Computação) – Programa de Pós-Graduação em Ciência da Computação, Universidade Federal de São Carlos, São Paulo, 2016. Disponível em: https://aberto.univem.edu.br/handle/11077/1483. Acesso em: 17 jul. 2020.

CHINO, D. Y. T.; ROMANI, L. A. S.; TRAINA, A. J. M. Construindo séries temporais de imagens de satélite para sumarização de dados climáticos e monitoramento de safras agrícolas. *Revista Eletrônica de Iniciação Científica,* [*s.l*], v. 10, n. 3, p. 1-20, 2010.

FELIPE, C. B. M.; FELIPE, A. A. C. *Análise do uso de metadados no auxílio à recuperação da informação em ambientes digitais*. [*s.l.*], 2016. Disponível em: http://www.liber.ufpe.br/home/wp-content/uploads/2016/09/04-Analise-no-uso-de-metadados_Felipe.pdf. Acesso em: 2 mar. 2020.

GONÇALVES, M. L. *et al.* Classificação não-supervisionada de imagens de sensores remotos utilizando redes neurais auto-organizáveis e métodos de agrupamentos hierárquicos. *Revista brasileira de cartografia,* [*s.l.*], v. 60, n. 1, abr. 2008. Disponível em: http://www.seer.ufu.br/index.php/revistabrasileiracartografia/article/view/44880. Acesso em: mar. 2021.

HALL, D.; JORDAN, J. *Human-centered information fusion.* [*S.l.*]: Artech House, 2010.

HAYKIN, S. *Redes neurais:* princípios e prática. 2. ed. Porto Alegre: Bookman, 2007.

JUAN, A.; TAULER, R. Data Fusion by Multivariate Curve Resolution. *In:* COCCHI, M. (ed.). *Data Handling in Science and Technology*. Netherlands: Elsevier, 2019. v. 31, cap. 8, p. 205-233.

MARCONI, M. A.; LAKATOS, E. M. *Metodologia do trabalho científico*. 8. ed. São Paulo: Atlas, 2017.

MOREIRA, F. C. *et al. Reconhecimento e classificação de padrões de imagens de núcleos de linfócitos do sangue periférico humano com a utilização de redes neurais artificiais*. 2002. Dissertação (mestrado) - Programa de Pós-Graduação em Ciência da Computação, Universidade Federal de Santa Catarina, Florianópolis, 2002. Disponível em: https://repositorio.ufsc.br/handle/123456789/82305. Acesso em: 17 jul. 2020.

NORMALIZED BURN RATIO. *NBR*. USGS, [2020]. Disponível em: https://www.usgs.gov/land-resources/nli/landsat/landsat-normalized-burn-ratio. Acesso em: 20 abr. 2020.

OLIVEIRA, R. A.; VITAL, L. P. Análise e indexação de imagens na rede Flickr. *Em Questão,* Porto Alegre, v. 21, n. 2, p. 7-30, maio/ago. 2015. DOI: http://dx.doi.org/10.19132/1808-5245212.7-30. Disponível em: https://seer.ufrgs.br/EmQuestao/article/view/50968/33977. Acesso em: 1 dez. 2019.

OSÓRIO, F. S.; BITTENCOURT, J. R. Sistemas Inteligentes baseados em redes neurais artificiais aplicados ao processamento de imagens. *In:* WORKSHOP DE INTELIGÊNCIA ARTIFICIAL, 1., 2000, Rio Grande do Sul. *Anais* [...]. Rio Grande do Sul: UNISC, 2000.

RIBEIRO, S. R. A.; CENTENO, J. S. Classificação do uso do solo utilizando redes neurais e o algoritmo MAXVER. *In:* SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO*,* 10., 2001, Foz do Iguaçu. *Anais* [...]. Foz do Iguaçu: INPE, 2001. p. 1341-1348. Disponível em: http://marte.sid.inpe.br/col/dpi.inpe.br/lise/2001/09.20.17.56/doc/1341.1348.074.pdf. Acesso em: 17 abr. 2020.

SIMIONATO, A. C. *Representação, acesso, uso e reuso da imagem digital.* 142 f. 2012. Dissertação (Mestrado em Ciência da Informação) – Programa de Pós-Graduação em Ciência da Informação da Faculdade de Filosofia e Ciências, Universidade Estadual Paulista Júlio de Mesquita Filho, Marília, São Paulo, 2012. Disponível em: https://www.marilia.unesp.br/Home/Pos-Graduacao/CienciadaInformacao/Dissertacoes/Simionato%20A.C._mestrado_C.I._2012.pdf. Acesso em: mar. 2021.

SRIVASTAVA, P. K. *et al.* Deriving forest fire probability maps from the fusion of visible/infrared satellite data and geospatial data mining. *Modeling Earth Systems and Environment,* [*s.l.*], v. 5, n. 2, p. 627-643, 2019. Disponível em: https://link.springer.com/article/10.1007/s40808-018-0555-5. Acesso em: mar. 2021.

VENKATESH, V. *et al*. Precision centric framework for activity recognition using Dempster Shaffer theory and information fusion algorithm in smart environment. *Journal of Intelligent & Fuzzy Systems,* [*s.l.*], v. 36, n. 3, p. 2117-2124, 2019. DOI: 10.3233/JIFS-169923.

YAMANE, G. A. C.; CASTRO, F. F. O estudo e a identificação dos padrões de metadados para a representação e a recuperação da imagem digital na perspectiva da web. *Em Questão,* Porto Alegre, v. 24, n. 1, p. 145-173, 2018. DOI: https://doi.org/10.19132/1808-5245241.145-173. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/71475. Acesso em: mar. 2021.

ZANOTTA, D. C.; FERREIRA, M. P.; ZORTEA, M. *Processamento de imagens de satélite*. São Paulo: Oficina de Textos, 2019.

Medição da informação científica na Web 2.0: explorando as possibilidades e limitações da plataforma Altmetric

Janinne Barcelos
Doutoranda, Universidade de Brasília (UnB), Brasília, Distrito Federal, Brasil.
Pesquisadora, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, Distrito Federal, Brasil.
URL http://lattes.cnpq.br/7729780084365345
E-mail: janbarcelos@hotmail.com

Diego José Macêdo
Mestre, Universidade de Brasília (UnB), Brasília, Distrito Federal, Brasil.
Tecnologista, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, Distrito Federal, Brasil.
URL http://lattes.cnpq.br/2205539000237712
E-mail: diegojmacedo@gmail.com

João de Melo Maricato
Doutor, Universidade de São Paulo (USP), São Paulo, Brasil.
Professor, Universidade de Brasília (UnB), Brasília, Distrito Federal, Brasil.
URL http://lattes.cnpq.br/3991129099537472
E-mail: diegojmacedo@gmail.com

Submetido em: 15/03/2021. Aprovado em: 23/06/2021. Publicado em: 28/07/2021 .

RESUMO

Com a evolução da Internet surgem novas formas de criação e compartilhamento de conhecimentos colaborativamente, culminando na Web 2.0 (Web social). Os produtos resultantes das atividades acadêmicas e científicas também passaram a circular nas mídias, despertando o interesse de pesquisadores na medição e no estudo das ações e interações relacionadas a tais atividades, surgindo a altmetria. Desde então, tecnologias, ferramentas e plataformas têm sido criadas para mensurar esses fenômenos, sendo, atualmente, a principal plataforma de coleta de dados altmétricos, a Altmetric, da empresa Digital Science. A presente pesquisa tem o objetivo explorar as possibilidades, limitações e características estruturais da Altmetric como fonte de dados e de indicadores altmétricos. A plataforma é analisada de maneira exploratória e experimental, a partir de dados da coleção SciELO e de entrevista com a diretora de relações de pesquisa da Altmetric, Stacy Konkiel. Como resultados, foram apresentadas e discutidas diversas possibilidades, limitações e características relacionadas aos fundamentos conceituais e à produção dos indicadores da plataforma, assim como os tipos de fontes/dados rastreados; amplitude de cobertura; cálculo do Altmetric score; popularidade e cobertura das mídias; possibilidades de estudo de correlações entre citações e menções.

Palavras-chave: Estudos métricos da informação. Mídiassociais. Web social. Ferramentas altmétricas. Altmetria.

Measuring scientific information on the web 2.0: exploring Altmetric's platafform possibilities and limitations

ABSTRACT

With the evolution of the Internet, new ways of collaboratively creating and sharing knowledge emerged, culminating in the Web 2.0 (Web social). With the evolution of the Internet, emerge new ways of creating and sharing knowledge collaboratively, culminating in the Web 2.0 (Web social). The products resulted from academic and scientific activities also began to circulate in the social media, inspiring the interest of researchers in the measurement and study of actions and interactions related to such activities, arising altmetrics. Since then, technologies, tools and platforms have been created to measure these phenomena, and Altmetric, from Digital Science, is currently the main platform for collecting altmetrics data. This research aims to explore the possibilities, limitations and structural characteristics of Altmetric as a source of data and altmetric indicators. The platform is analyzed in an exploratory and experimental way, based on data from SciELO's collection and an interview with Altmetric's director of research relations, Stacy Konkiel. As a result, several possibilities, limitations and characteristics related to the conceptual foundations and the production of the platform indicators were presented and discussed, as well as the types of sources / data tracked; coverage range; calculation of the Altmetric score; popularity and media coverage; possibilities of studying correlations between citations and mentions.

Keywords*: Informetrics. Social media. Social web. Altmetric tools. Altmetrics.*

Medición de información científica en la Web 2.0: explorar las posibilidades y limitaciones de la plataforma Altmetric

RESUMEN

Con la evolución de Internet surgen nuevas formas colaborativas de crear e intercambiar conocimientos, que culminan en la Web 2.0 (Web Social). Los produtos resultantes de las actividades académicas y científicas también comenzaron a circular en los medios de comunicación despertando el interés de investigadores en la medición y estudios de las acciones e interacciones relacionadas a dichas actividades, surgiendo de esta forma, la altmetría. Desde entonces, tecnologías, herramientas y plataformas han sido creadas para medir estos fenómenos, entre ellas se encuentra Altmetric, de la empresa Digital Science, principal plataforma de coleta de dados altmétricos. Esta investigación tiene como objetivo explorar las posibilidades, limitaciones y características estructurales de Altmetric como fuente de datos e indicadores altmétricos. La plataforma es analizada de forma exploratoria y experimental, a partir de datos de la colección Scielo y entrevista con la directora de investigación de Altmetric, Stacy Konkiel. Como resultado fueron planteadas y discutidas diversas posibilidades, limitaciones y características relacionadas a los fundamentos conceptuales y a la producción de indicadores de la plataforma, así como los tipos de fuentes/datos rastreados; rango de cobertura, cálculo de Almetric score; popularidad y cobertura mediática; y posibilidades de estudios correlacionales entre citaciones y menciones.

Palabras clave: *Estudios de información métrica. Redes sociales. Web social. Herramientas altmétricas.*

INTRODUÇÃO

Principalmente a partir do início dos anos 2000, quando a Internet começou a oferecer a possibilidade de compartilhamento e de criação de novos conhecimentos por meio de conteúdos colaborativos e mídias sociais, as dinâmicas de transmissão da informação se modificaram (LEMOS, 2002). A visada sociológica de Castells (1999) - que explica o papel das tecnologias da informação e a mudança de paradigma que elas representaram na virada do século XX - alerta para a web 2.0 (web social) como geradora de novos tipos de informação advindas das interações entre os usuários da rede. Diante do uso crescente das mídias sociais e da apropriação dessas mídias para a disseminação das pesquisas científicas, faz-se necessário ampliar as discussões sobre como medir o impacto social da produção acadêmica no ambiente virtual. Sobretudo, como avaliar a atenção que recebem e o índice de visibilidade que atingem esses produtos de pesquisa.

Certamente, o estudo da comunicação científica na Internet não é um fenômeno novo. A aplicação de métodos bibliométricos contribuiu – através do exame de links, mailing lists e estruturas da rede acadêmica – para uma perspectiva complementar da tradicional análise de citações. "Entretanto, esta disciplina não foi capaz de superar certas limitações inerentes às metodologias, métodos e fontes de informação utilizadas" (TORRES-SALINAS; CABEZAS-CLAVIJO; JIMENEZ-CONTRERAS, 2013, p. 2, tradução nossa). Desse contexto, ainda emergiram áreas como cibermetria, webometria e outras inúmeras variantes com diferentes níveis de proximidade, igualmente apoiadas na premissa de que o impacto provocado por produções científicas fica evidente a partir da contagem de citações. Mas, da mesma maneira que o significado do que vem a ser fazer pesquisa tem mudado drasticamente com os avanços das tecnologias da informação, também mudaram as definições para o que constitui, de fato, um estudo de impacto (ROEMER; BORCHARDT, 2015).

Conforme explica a *National Information Standards Organization* (NISO), o comportamento do leitor online, o conteúdo das redes de interações e as referências em mídias sociais são todos importantes indicadores de impacto dos resultados da pesquisa que não estão refletidos nas contagens de citação (O'NEIL, 2016). Tendo em vista essas limitações e a expansão das mídias sociais, é necessário lançar mão de novas metodologias para compreender a comunicação científica e o fenômeno complexo que é a relação Ciência, Tecnologia e Sociedade. Este é um dos principais problemas endereçados pela altmetria e pelo uso de seus indicadores (PRIEM; HEMMINGER, 2010; NASSI-CALÒ, 2015; NISO, 2016).

A altmetria, definida como um tipo de estudo métrico da informação para a análise da atividade acadêmica baseada na Web 2.0, está baseada na ideia de que seus indicadores fornecem uma visão complementar da relevância e do impacto dos produtos de resultantes das atividades de pesquisa. Campo recente na Ciência da Informação, em plena fase de construção, a Altmetria e seus indicadores estão sendo cada vez mais usados e discutidos como uma expansão das ferramentas disponíveis para medir o impacto da pesquisa. Os pesquisadores que apoiam seu uso afirmam que os indicadores altmétricos têm o potencial de complementar ou melhorar os sistemas de avaliação científica mais tradicionais (PRIEM *et al.*, 2010; PIWOWAR, 2013; BARROS, 2015).

Para seus defensores, os indicadores altmétricos possibilitam analisar o impacto social da ciência "em canais de mídia social que são mais inclusivos e democráticos do que os editores e os bancos de dados de citação [com] potencial para reverter décadas de marginalização no sistema atual" (ALPERIN, 2013, *online*, tradução nossa). Por outro lado, "como todo novo conceito, costuma gerar dúvidas e questões sobre sua legitimidade, principalmente pelo fato de utilizar ferramentas 'informais' para medir o impacto da ciência, essencialmente formal" (NASSI-CALÒ, 2017, *online*).

O que todos parecem concordar, desde a proposição de Priem *et al.* (2010) em seu manifesto, é que a Altmetria está em sua fase inicial, com muitas perguntas ainda sem respostas. Não há consenso sobre o uso dessas métricas na academia, contudo, "dada a crise enfrentada pelos filtros existentes e a rápida evolução da comunicação científica, a velocidade, a riqueza e amplitude dos indicadores altmétricos fazem valer o investimento nesta iniciativa" (PRIEM *et al.*, 2010, *online*, tradução nossa). Razão pela qual despontaram inúmeras ferramentas altmétricas de diferentes organizações com diferentes diretrizes para coleta e geração de relatórios de dados.

Atualmente, entre os provedores de dados altmétricos com cobertura mais abrangente está a plataforma Altmetric (ROBINSON-GARCIA *et al.*, 2014). Disponível desde o início de 2012, a Altmetric monitora fontes não tradicionais como Facebook, Twitter, blogs, sites de notícias e outras mídias sociais, buscando links e referências de estudos acadêmicos publicados. Até o momento, seu banco de dados contém menções de mais de 9 milhões de resultados de pesquisas - incluindo livros, artigos de periódicos, *datasets*, *preprints* e relatórios. Ou seja, a Altmetric trata-se de uma plataforma com larga diversidade de novos dados e potenciais de pesquisa. Contudo, pouco tem sido o debate sobre a amplitude desses dados, como interpretá-los e que alternativas de pesquisa esta plataforma pode oferecer (ROBINSON-GARCIA *et al.*, 2014).

Inspirado nas pesquisas de Robinson-Garcia *et al.* (2014), este estudo tem por objetivo explorar as possibilidades, limitações e características estruturais da Altmetric como fonte de dados e de indicadores altmétricos. Por meio de análise exploratória da plataforma, de experimento com dados da coleção SciELO e de entrevista com a diretora de relações de pesquisa da Altmetric, Stacy Konkiel, analisa-se especificamente: as bases conceituais que orientam a produção dos indicadores; tipos de fontes/dados rastreados; amplitude de cobertura; cálculo do *Altmetric score*; mídias que apresentam mais menções (mais populares) e possibilidades de estudo de correlações entre citações e menções na plataforma.

METODOLOGIA

Trata-se de um estudo exploratório, descritivo e analítico, cujo objetivo é identificar e examinar características de usabilidade da plataforma Altmetric, de maneira qualiquantitativa. Para tanto, a metodologia foi dividida em duas etapas. A primeira delas compreendeu uma entrevista estruturada com a diretora de relações de pesquisa da Altmetric, Stacy Konkiel, e análise de conteúdo do website. Esses procedimentos iniciais propiciaram a identificação de conceitos teóricos sobre altmetria e o impacto social que orienta a produção dos indicadores na Altmetric, assim como a identificação de questões práticas como os tipos de fontes/dados, características das fontes rastreadas, como as fontes são selecionadas e quais são os identificadores persistentes utilizados. O roteiro da entrevista, esboçado no quadro 1, foi baseado em discussões realizadas durante reunião aberta do Grupo de Pesquisa xxxxxxxxxxxxxxx da Universidade xxxxxxxxxxxxxxxxxxxxxxxxxxx, no dia 26 de outubro de 2019.

Quadro 1 − Roteiro de entrevista estruturada com a diretora de relações de pesquisa da Altmetric, Stacy Konkiel.

MOTIVAÇÃO	PERGUNTAS
Visão e conceitos sobre altmetria	O que é altmetria e o que ela mede? Para que a altmetria é útil?
Visão e conceitos sobre impacto	O que significa 'impacto social da pesquisa' em termos de altmetria? Como está linkada a ideia de impacto entre a plataforma Altmetric e Dimensions?
Questões práticas	AAltmetric pode ser considerada uma base de dados científica? O que ela cobre? Como são selecionadas as fontes indexadas na plataforma Altmetric? Como o Altmetric score é calculado? Como podemos obter dados altmétricos sem um DOI?

Fonte: Dados da pesquisa, 2019.

Na segunda etapa, realizada no dia 05 de dezembro de 2019, explorou-se o comportamento e o potencial técnico da Altmetric por meio de experimentos de busca combinada. No campo "nome do editor", buscou-se pelo termo "Scielo" e, no campo "tipo de output", selecionou-se "artigos". No total, foram recuperados 81.622 artigos, em relatório fornecido pelo Altmetric Explorer. A partir desse conjunto de dados, foi possível verificar a amplitude da base; analisar a recuperação dos dados; avaliar como é feito o cálculo do *Altmetric score*; identificar mídias que apresentam mais menções/mais populares e testar possibilidades de estudo de correlações entre citações e menções a partir da Altmetric.

Convém destacar que este estudo se utiliza do conjunto de referências da SciELO para estudar a Altmetric, e não o contrário. Logo, não serão feitas quaisquer inferências sobre a popularidade ou a qualidade dos artigos indexados pela SciELO. A escolha da Scientific Electronic Library Online (SciELO) foi aleatória. Para os fins deste trabalho, utilizou-se a versão gratuita do Altmetric Explorer, cujo acesso é permitido para bibliotecários e pesquisadores, desde que um acordo de compartilhamento dos dados seja assinado e respeitado. Vinculada à Digital Science, a Altmetric pode ser acessada no sítio www.altmetric.com.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

A partir da entrevista realizada com a diretora de relações de pesquisa, Stacy Konkiel, ficou evidente que as atividades geridas na plataforma Altmetric são norteadas por três conceitos basais. O primeiro diz respeito à definição de altmetria, que é entendida como "métricas e dados qualitativos da web que refletem a atenção e o engajamento que a pesquisa recebe em jornais, documentos de políticas, patentes, mídias sociais, blogs e outras fontes online" (KONKIEL, 2019). O segundo, refere-se à utilidade da altmetria: ajudar o público interessado a "entender 'quem' e 'o que' está sendo falado sobre a pesquisa no ambiente online" (KONKIEL, 2019).

O último estende-se ao significado de impacto, entendido como "as mudanças que a pesquisa provoca nas percepções do público e se ela está fazendo uma diferença positiva na vida das pessoas" (KONKIEL, 2019). Dito de outra forma, trata-se de uma compreensão mais holística de impacto que envolve, além das citações, o alcance da pesquisa fora dos muros da academia. Nesse sentido, atualmente, a Altmetric busca fornecer, tanto registros de menções online encontrados para um item (em comunicações não formais), quanto dados de citações em publicações de periódicos, fornecidos pela plataforma Dimensions (produto da Digital Science que se propõe a rastrear dados de pesquisas desde o financiamento até sua publicação).

Da análise de conteúdo da Altmetric, verificaram-se tipos distintos de informações que podem ser agrupados em quatro categorias: 1) pontuações de cada publicação, incluindo score e quantidade de menções em cada uma das mídias; 2) informações sobre o periódico, que abarcam o nome, o número do ISSN e status de Open Access (OA); 3) descrição do artigo recuperado, que conta com referência bibliográfica, área do conhecimento à qual está vinculado, data de publicação do artigo, data de inclusão na Altmetric, URL e número do DOI; 4) contagem do total de citações e de referências das três últimas publicações em que o artigo aparece citado em dados indexados pela Dimensions.

De acordo com Konkiel (2019), a Altmetric "não é um banco de dados tradicional". Ela não busca indexar as pesquisas já publicadas, mas as pesquisas mencionadas nas fontes não tradicionais rastreadas pela base (Facebook, Twitter, LinkedIn, IIFI CLAIMS, cerca de nove mil blogs, mais dois mil sites de notícias, etc). Dessa forma, as métricas são calculadas com base no número de vezes em que uma produção científica foi clicada, lida, compartilhada, mencionada, discutida ou revisada por usuários das mídias indexadas pela plataforma (KONKIEL, 2019).

Via de regra, o rastreamento dos dados é feito por meio de 9 códigos ou identificadores persistentes: 1) PubMedID (normalmente associado à pesquisa em ciências da saúde); 2) arXiv ID (física, matemática e ciências da computação); 3) ADS ID (dados de astrofísica); 4) SSRN ID (ciências sociais); 5) RePEC ID (pesquisa econômica); 6) Handle.net (repositórios institucionais); 7) URN (Nome Uniforme do Recurso); 8) ISBNs e 9) DOIs. A única exceção do rastreio automatizado acontece em blogs, notícias e documentos de Políticas Públicas (PPs), em que são usadas técnicas de mineração de texto e curadoria manual (somente em inglês) em uma lista de feeds RSS. Neste caso, a equipe responsável procura por hiperlinks e referências a artigos acadêmicos, a periódicos e a autores no conteúdo de reportagens e posts de blogs (ALTMETRIC, 2019).

No que diz respeito às ferramentas de busca, a Altmetric fornece estratégias de buscas simples e avançadas, com opções de filtros que possibilitam cruzamento de dados. Na busca geral, é possível pesquisar por título, palavras-chave, autor, editora, nome do periódico e agência financiadora. Na busca avançada, pode-se combinar filtros entre tipos de documentos (artigo, livro, capítulo de livro, conjunto de informações, registros médicos e notícias). Também é possível adicionar identificadores acadêmicos, caso existam, como por exemplo, DOI, ISBN, PubMed, ID, arXiv e outros; filtrar por periódico específico ou lista determinada de periódicos e estabelecer argumentos de datas (qualquer tempo, último dia, últimos três dias, última semana, último mês, últimos três meses, últimos seis meses e último ano). Contudo, ainda não é possível pesquisar por data específica, o que inviabiliza a reprodução dos resultados de uma busca, dada a volatilidade com que as mídias digitais se atualizam.

No que diz respeito às fontes da Altmetric, foram identificadas 17 mídias, de diferentes tipologias. As métricas referentes a cada mídia estão detalhadas no *donut* e podem ser baixadas pelo usuário em formato CSV (Comma-Separated Values). De acordo com Konkiel (2019), a Altmetric seleciona quais fontes rastrear com base em vários fatores, incluindo sua popularidade, a frequência com que pesquisas são mencionadas, a demanda pelos dados e se é tecnicamente viável coletá-los. No quadro 2, encontram-se os nomes das mídias identificadas, sua descrição, os campos de dados que são recuperados pela base em CSV e o peso estipulado pela Altmetric para cada mídia no cálculo do *score*. É importante observar que as citações da Dimensions, os marcadores no CiteULike e os leitores do Mendeley não estão incluídos neste cálculo.

Quadro 2 − Identificação das mídias, dados recuperados e peso por mídia na Altmetric

MÍDIAS	TIPOS	DADOS RECUPERADOS	PESO NO SCORE
Notícias	Sites de divulgação	Título da notícia; URL da notícia; data e hora de publicação; licença; resumo; nome da mídia; URL da mídia; ID da mídia; imagem da mídia	8
Blogs	Lista de RSS com curadoria manual	Título do Blog; título do post; URL do post; data e hora de publicação; resumo; nome do autor; URL do autor; descrição do autor	5
Políticas	Sites que contém documentos de PPs	Título do documento; autor do documento; capa do documento; descrição do autor; link direto para o documento	3
Patentes	Sites de escritórios internacionais de patentes	Tipo de menção; data da menção; título da menção; país; ID da menção; título da publicação; área de conhecimento da publicação; URL da publicação	3

(Continua)

Quadro 2 – Identificação das mídias, dados recuperados e peso por mídia na Altmetric (Conclusão)

MÍDIAS	TIPOS	DADOS RECUPERADOS	PESO NO SCORE
Twitter	Microblogging	URL; data e hora de publicação; licença; resumo; nome do autor nome; imagem do autor; número de seguidores; ID do tweet; tipo de público; país	1
Peer Review	Sites de revisão por pares pós-publicação	Informações das plataformas especializadas em revisão por pares Publons e PubPeer; comentários postados nas duas plataformas, data de postagem, autor da postagem.	1
Weibo	Rede Social	Para visualizar informações da Weibo, uma rede social chinesa, é preciso se cadastrar e ser um cidadão chinês ou ser um funcionário de governo. Não rastreável desde 2015, mas dados históricos mantidos.	1
Facebook	Rede Social	Título da menção; URL da menção; data e hora de publicação; resumo; nome do autor; URL do autor; nome da página do Facebook; imagem do autor; ID do autor	0.25
Wikipedia	Enciclopédia colaborativa	Verbetes da enciclopédia em que o artigo é citado, autor do post, data do post, idioma do post, pequena descrição do post, link para o post na plataforma original	3
Google+	Rede Social	Título da menção; URL da menção; data e hora de publicação; resumo; nome do autor; URL do autor; imagem do autor; ID do autor. Não rastreável desde 2019, mas dados históricos mantidos.	1
Linkedin	Rede Social Profissional	Total de usuários únicos; nome dos usuários; total de posts; título do post; resumo; data e hora de publicação; nome do autor; descrição do autor; URL do post; URL do logotipo do grupo; nome do grupo; descrição do grupo. Não rastreável desde 2014, mas dados históricos mantidos.	0,5
Reddit	Site de notícias	Título da notícia; URL reddit; data e hora de publicação; nome do autor; URL do autor; identificação do autor; seguidores	0,25
Pinterest	Rede Social	Menção de URL; imagem da menção; data e hora de publicação; resumo; nome do autor; mural. Não rastreável desde 2013, mas dados históricos mantidos.	0.25
F1000	Site de divulgação pós-publicação	Recomendação do F1000; data de publicação (provavelmente da última atualização); tipo de recomendação	1
Q&A	Sites de perguntas e respostas (Stack Overflow)	Título; URL de discussão; data e hora de publicação; resumo; ID do autor	0,25
Youtube	Site de compartilhamento de vídeos	Título do video; URL do vídeo; imagem de vídeo; data e hora de publicação; licença; resumo; ID do YouTube; nome do autor; ID do autor	0,25
Syllabi	Coletor de dados sobre livros publicados em sites de universidades	Nome da instituição; área de conhecimento vinculada à publicação	1

Fonte: Adaptado de Altmetric (2019).

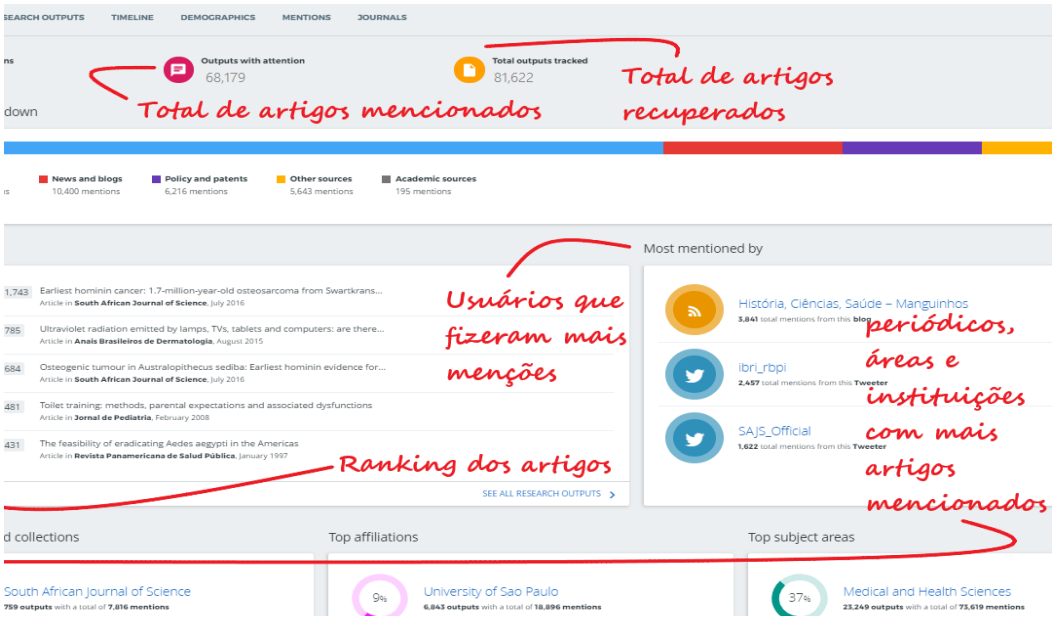
AMPLITUDE DE COBERTURA, RECUPERAÇÃO DE DADOS, CÁLCULO DO *ALTMETRIC SCORE* E *DONUT*

A partir da amostra extraída sobre os outputs da SciELO, verificou-se que os 81.622 artigos recuperados pela Altmetric contabilizaram um total de 198.385 menções entre 17 tipos de mídias sociais diferentes, marcadores do CiteULike, *downloads* do Mendeley e citações da Dimensions. A recuperação desses dados demonstra que a plataforma tem o potencial de extrair dados das fontes e das mídias a que se propõe. Contudo, conhecendo-se a quantidade total de artigos da coleção SciELO, que, em dezembro de 2019, era de 402.318, pode-se constatar que a amplitude da cobertura da Altmetric chega a 20% (81.622 artigos recuperados no contexto da Scielo). Nota-se que Robinson-Garcia *et al.* (2014), ao estudarem a cobertura dos artigos da base de dados Web of Science na Altmetric, chegaram a um resultado muito próximo (19%).

Para apresentar os dados recuperados, a Altmetric fornece seis áreas distintas, denominadas "destaques", "resultados da pesquisa", "linha do tempo", "demografia", "menções e "periódicos". Na área "destaques" (figura 1), é agrupado o total de menções/por mídia de todos os artigos e são apresentados rankings com indicadores variados (artigos com maiores scores; usuários que mais mencionaram os artigos; periódicos, disciplinas e afiliações que tiveram mais artigos mencionados), além das últimas menções e da atenção por região.

Na área "resultados de pesquisa", o Altmetric Explorer fornece, para cada artigo, um *score* e um *donut* que representa, por meio de cores 'quanto' e 'qual tipo de atenção', um resultado de pesquisa recebeu. Segundo a Altmetric (2019), o cálculo do *score* é derivado de um algoritmo automatizado e representa uma razão entre 3 fatores: 1) volume de menções - a pontuação do artigo aumenta na medida em que mais pessoas o mencionam. A pontuação só é creditada por pessoa em cada fonte, ou seja, o *score* não considera menções repetidas por usuário; 2) autoria das menções - a frequência com que um autor fala sobre os artigos e para que público ele fala interferem no resultado do *score*; 3) tipo de fonte - cada tipo de fonte influencia de maneira distinta a pontuação final, como esboçado no quadro 2.

Figura 1 – Área de destaques do Altmetric Explorer



Fonte: Captura de tela da www.altmetric.com (2019).

Contudo, a Altmetric falha em dizer como, **exatamente**, o cálculo do *score* é feito. Nos testes aplicados ao conjunto de dados da SciELO, notou-se que o cálculo não é uma "somatória simples" dos pesos por tipos de mídias (listados no quadro 2). Para exemplificar, pode-se tomar como modelo o artigo com maior *score* do universo recuperado, identificado neste estudo como Artigo 1. Segundo o relatório fornecido pelo Altmetric Explorer, no dia 05 de dezembro de 2019, o *score* do Artigo 1 era 1.743, sendo mencionado 274 vezes em sites de notícias, 17 vezes em blogs, 181 no Twitter, 29 no Facebook, duas na Wikipédia, oito no Google+ e uma no Reddit. Ao somar os pesos de cada tipo de mídia em que o Artigo 1 foi mencionado (considerando os pesos estipulados pela própria Altmetric no quadro 2), chega-se a um total diferente do *score* apresentado pela base: 2.479,50 (tabela 1).

Tabela 1 – Somatória dos pesos por tipo de mídia do Artigo 1, segundo valores publicados pela Altmetric, em 2019

TIPO DE MÍDIA	Nº DE MENÇÕES	PESO POR MÍDIA	SOMA
Notícias	274	8	2.192
Blog	17	5	85
Twitter	181	1	181
Facebook	29	0,25	7,25
Wikipedia	2	3	6
Google+	8	1	8
Reddit	1	0,25	0,25
TOTAL			2.479,50

Fonte: Dados de pesquisa (2019)

Em seu website, a Altmetric (2019) argumenta que o *score* sempre deve ser um número inteiro. Portanto, as menções que contribuem com pesos menores que 1, são arredondadas (para cima). Consequentemente, se for rastreada, apenas uma postagem sobre certo artigo no Facebook, esta será equivalente a 1 (1 x 0,25 = peso 1). Mas, se forem rastreadas mais três postagens no Facebook para o mesmo artigo, a pontuação ainda aumentaria apenas em 1 (4 x 0,25 = peso 1). Ainda segundo a Altmetric (2019), a pontuação dos artigos mencionados na Wikipedia é estática. Isto é, independentemente do número de verbetes que acumula a pontuação desse artigo aumentará apenas em 3.

No quis diz respeito às menções feitas em sites de notícias, documentos de políticas públicas e patentes, a Altmetric (2019) explica apenas que os pesos são pontuados por fonte, sem especificar, no entanto, quanto precisamente pesa cada fonte. Não foi encontrada no website ou no Altmetric Explorer qual é, de fato, a metodologia e a base de cálculos utilizada para dizer se este ou aquele documento deve ter *score* x ou y. Esta falta de clareza no procedimento impossibilita tanto a auditoria do índice, quanto a solução de problemas rotineiros para o marketing científico, como: quais seriam os canais de notícias mais populares? Em que canais de divulgação as revistas deveriam investir mais? Quando questionada sobre o assunto, Stacy Konkiel informa que a pontuação de notícias varia de acordo com a fonte, sendo que as fontes da camada 1 valem 8 pontos, as da camada 2 valem cinco pontos e as da camada 3 valem três. Mas concorda que, atualmente, "uma análise detalhada do motivo pelo qual cada trabalho é pontuado e como é pontuado não é possível dadas as nuances do sistema de pontuação. Sabemos que isso não é o ideal e estamos pensando em maneiras de tornar nossos mecanismos de pontuação mais transparentes".

Para além do *Altmetric score*, ao clicar no *donut*, o usuário é direcionado para a página de detalhes do artigo (figura 2), na qual se encontra a descrição dos itens que contribuíram para o score, agregados por mídia social. Há também uma exibição demográfica que contém os países e o tipo de público responsável pelas menções. Esta informação está baseada nas contas de usuários em que é possível identificar o país de origem dos usuários que mencionam o documento (ou seja, quando a origem é declarada pelo usuário). A partir destas métricas, é possível tabular e comparar informações que ajudam a compreender como a pesquisa tem sido recebida e interpretada.

Entre as perguntas que podem ser respondidas com estes tipos de indicadores estão: Esta publicação vem sendo noticiada/divulgada? O que estão falando sobre a pesquisa? Em que países estão falando sobre a pesquisa? Em quais mídias e quantas vezes em cada uma delas a publicação foi mencionada?

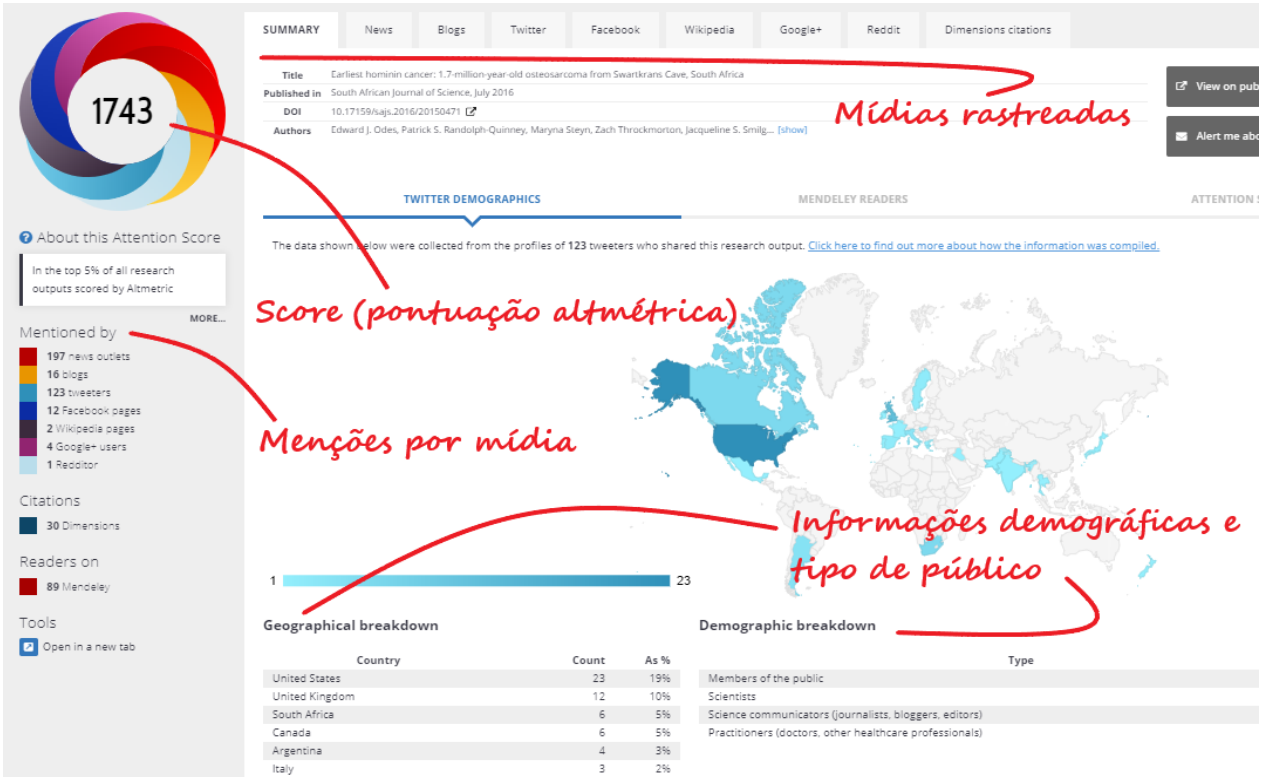
Durante a análise dos dados extraídos da área "resultados de pesquisa", notaram-se diferenças entre a separação dos registros recuperados em CVS e aqueles exibidos no donut (por exemplo, *tweets* e *retweets*). Fato também observado no estudo de Robinson-Garcia *et al.* (2014). Foram constatados ainda, erros ocasionais da base em relação à data da publicação dos artigos e à data de sua menção nas mídias rastreadas. Observaram-se alguns casos em que as menções de blogs e de sites de notícias apareceram antes mesmo da publicação do artigo no periódico. Provavelmente, isso acontece com os itens inicialmente disponibilizados como preprints, que, depois de formalmente veiculados em periódicos, têm sua data de publicação atualizada. Outras possibilidades são revistas que antecipam a publicação de fascículos ou que divulgam *online* as versões preliminares dos artigos conforme forem sendo aprovados. Porém, seria necessária pesquisa mais aprofundada para verificar se esse é um erro de processamento dos dados pela plataforma ou se se tratam das possibilidades apresentadas. De qualquer forma, ressalta-se a necessidade de atenção do pesquisador na consolidação dos dados e, eventualmente, a correção manual de incoerências detectadas.

Ademais, observaram-se casos em que os links de acesso aos blogs e aos sites de notícias estão indisponíveis. Entre os exemplos que podem ser citados, está o artigo intitulado *The remarkable journey of adaptation of the Plasmodium falciparum malaria parasite to New World anopheline mosquitoes*, que teve todos os links das notícias rastreadas quebrados (a página de detalhes do artigo está disponível em: https://www.altmetric.com/details/3742834). Acredita-se que isso aconteça em razão da altmetric recorrer a mecanismos de agregação (como o Moreover.com) para rastreio e armazenamento dos sites e blogs que referenciam artigos acadêmicos, periódicos e autores. Entretanto, o conteúdo recuperado pode ser acessado diretamente no site ou blog em que o output foi mencionado.

Nas áreas "linha do tempo", "demografia" e "menções", foi possível estabelecer relações sobre quando, onde e quem está falando sobre cada artigo do universo sob análise (81.622 artigos da SciELO). Juntas, as três áreas oferecem listas de componentes sociais, geográficos e conteudísticos que possibilitam responder: onde a pesquisa está sendo comentada no momento? Quais mensagens estão sendo divulgadas sobre a pesquisa? Que tipo de usuário está dizendo algo sobre a pesquisa? etc.

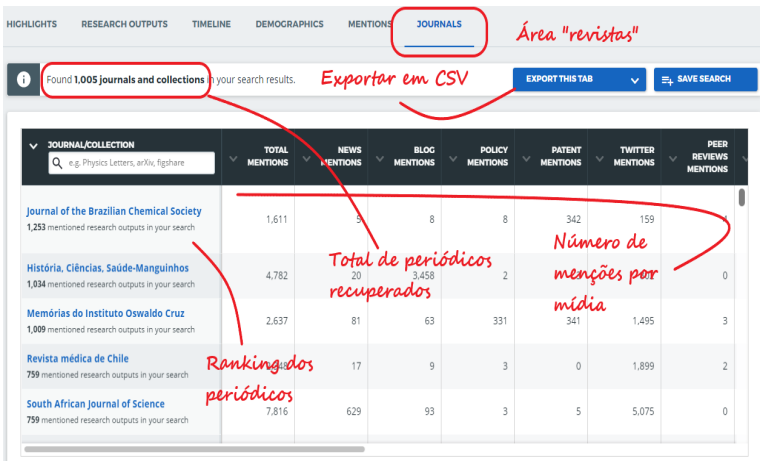
Tais indicadores também ajudam a medir o sucesso das atividades de divulgação empreendidas pelos autores, periódicos e instituições, identificar públicos interessados e guiar estratégias futuras. Porém, para isso, será exigido do pesquisador a capacidade de estabelecer conexões e divergências entre os indicadores, bem como conhecimento prévio de técnicas de análise de dados, estatística e análise de conteúdo. Na área "revistas", os números de menções estão aglomerados por periódico em um ranking similar aos indicadores oferecidos nas bases de dados tradicionais (Figura 3).

Figura 2 – Página de detalhes de um artigo no Altmetric Explorer



Fonte: Captura de tela da www.altmetric.com (2019).

Figura 3 - Área “revistas” do Altmetric Explorer



Fonte: Captura de tela da www.altmetric.com (2019).

MÍDIAS MAIS POPULARES E TWITTER VESUS MENDELEY

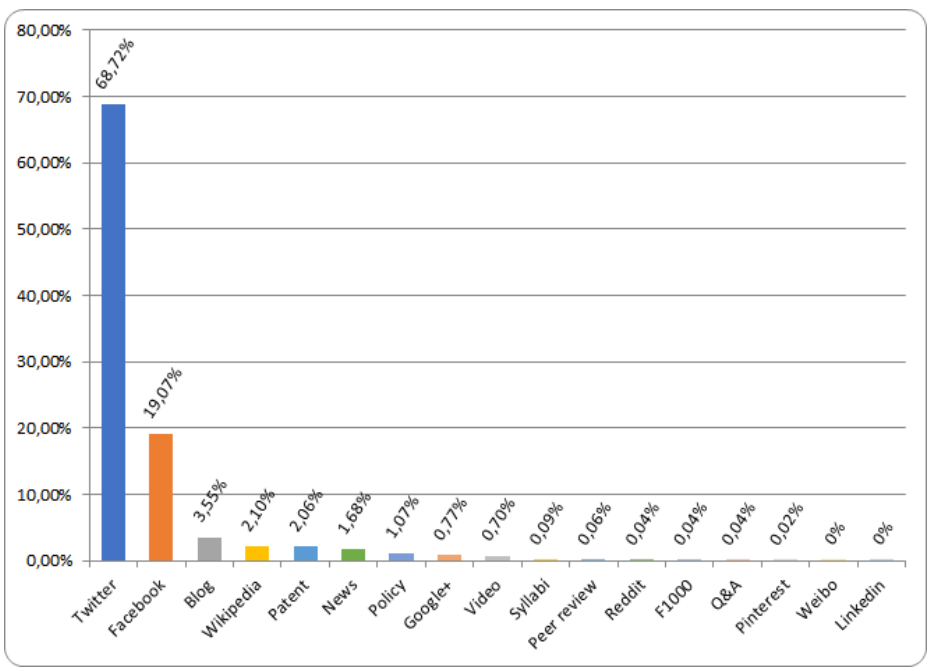
Do conjunto de dados extraídos sobre os artigos da SciELO, também buscou-se verificar que mídias acumulam mais menções, podendo-se deduzir quais são as mais populares. No total, a Altmetric recuperou 81.697 artigos publicados e indexados na SciELO. Destes, 83,46% foram mencionados pelo menos uma vez nas mídias rastreadas pela Altmetric. Ao verificar cada mídia, observou-se que o Twitter (68,72%) e o Facebook (19,07%) apresentam o maior número de menções, sugerindo certa popularidade entre os usuários (Gráfico 1). Este mesmo fato também foi observado nos estudos de Hughes *et al.* (2012) e de Ahmed (2017).

É importante destacar que o número de vezes que os artigos foram salvos por leitores do Mendeley não estão incluídos no Gráfico 1 pelo fato de não computarem no momento de calcular o *Altmetric score* e por terem números muito maiores quando comparado com as demais mídias e fontes de dados altmétricos cobertos pela Altmetric.

O número de downloads dos artigos da SciELO pelos leitores do Mendeley foi de 1.782.275, totalizando aproximadamente 90% do total, sendo disparada a mídia mais utilizada para provimento de dados altmétricos.

Ao incluir os dados do Mendeley nas análises apresentadas no Gráfico 1, observa-se que o Twitter cairia drasticamente para 6,9% na popularidade (contra os 90% do Mendeley). Quando se compara esses dados com os do estudo de Robinson-Garcia *et al.* (2014) observa-se que, apesar dos dados serem de bases diferentes, há fortes indícios de que houveram mudanças drásticas no cenário desde então. Os autores observaram, na ocasião, que o Twitter era a fonte que contava com mais dados altmétricos (87,1%) seguidos por Mendeley (64,8%). Esses resultados demonstram que a importância do Mendeley vem aumentando fortemente para o oferecimento de dados altmétricos, podendo haver relação com o aumento de usuários da ferramenta. Outra possibilidade poderia ser as mudanças nas participações nacionais dos leitores de Mendeley, que têm aumentado em alguns países e diminuído em outros (FAIRCLOUGH; THELWALL, 2015).

Gráfico 1 – Acúmulo de menções por mídia de artigos da coleção SciELO



Fonte: Dados da pesquisa (2019).

No entanto, adverte-se que é preciso cuidado ao inferir comparações entre mídias, já que suas políticas de dados são distintas. Embora o Twitter apresente número consideravelmente maior de menções que o Facebook, é importante notar que talvez isso aconteça porque a Altmetric rastreia apenas mensagens de perfis públicos da rede social - excluindo, por exemplo, menções de indivíduos e grupos com perfis fechados. Dessa forma, utilizando-se apenas de evidências fornecidas pela Altmetric, não é possível afirmar se esta ou aquela mídia é mais popular entre um determinado grupo. Se esse fosse um dos objetivos desta pesquisa, seria preciso relacionar os dados recuperados pela Altmetric a indicadores externos e/ou a outras pesquisas.

Ainda sobre a popularidade das mídias, argumenta-se que blogs (3,55%), Wikipédia (2,10%), patentes (2,06%) e notícias (1,68%) apresentam números significativos quando se considera que plataformas desse tipo contam com conteúdo mais extenso e, portanto, com processos de maturação mais demorados do que mensagens no Twitter, por exemplo, que se limitam a até 280 caracteres para cada postagem.

POSSIBILIDADE DE ESTUDO DE CORRELAÇÕES ENTRE CITAÇÕES E MENÇÕES NA ALTMETRIC

Para verificar se é possível estabelecer correlações entre indicadores altmétricos (menções) e indicadores bibliométricos (citações), utilizando dados da plataforma Altmetric, foram extraídos os dados da coleção SciELO. A plataforma possibilitou a extração de todos os registros desejados, sem limitações ou imposição de extração apenas de amostras. Assim, extraiu-se o universo desejado (81.622), que se refere ao total de artigos da coleção SciELO.

Os registros podem ser extraídos no formato CSV e posteriormente analisados. Os metadados dos artigos da planilha exportada na plataforma Altmetric são bastante diversos.

Além do score altmetric, título do artigo, título da revista, número do ISSN, indicação se é Open Access, data de publicação, identificadores persistentes, disponibiliza indicadores das quantidades de menções em cada uma das mídias e das fontes de informação.

No que se refere a indicadores bibliométricos, a plataforma Altmetric disponibiliza o número de citações a partir dos dados da base de dados Dimensions. O número de citações de cada artigo pode ser facilmente correlacionado com o score altmetric, bem como com outras contagens de menções e ocorrências nas mídias cobertas pela plataforma. Realizando-se a análise da correlação entre o universo extraído da coleção SciELO, de 81.622 artigos, chega-se a uma correlação bem fraca de menos de 0,07 entre as citações e o *Altmetric score*. Outros estudos têm encontrado correlação baixa entre citação e indicadores altmétricos, a exemplo de Costas, Zahedi e Wouters (2014) que, ao estudar correlação entre dados da Altmetric com dados de citação da WoS, identificou correlações entre 0,15 e 0,93 aproximadamente, levemente maior que a encontrada na SciELO, porém bastante fraca também.

Com relação ao potencial da Altmetric para a realização de estudos de correlação entre citações e dados altmétricos, a plataforma poderia disponibilizar dados de outras fontes de dados de citações, tal como das bases de dados Web of Science, Scopus, SciELO e Crossref para que fosse possível realizar outras análises e correlações. No entanto, reconhece-se que algumas bases de dados e editores não permitem a utilização dos dados por terceiros e que a utilização de dados de citação pode não depender unicamente da Altmetric.

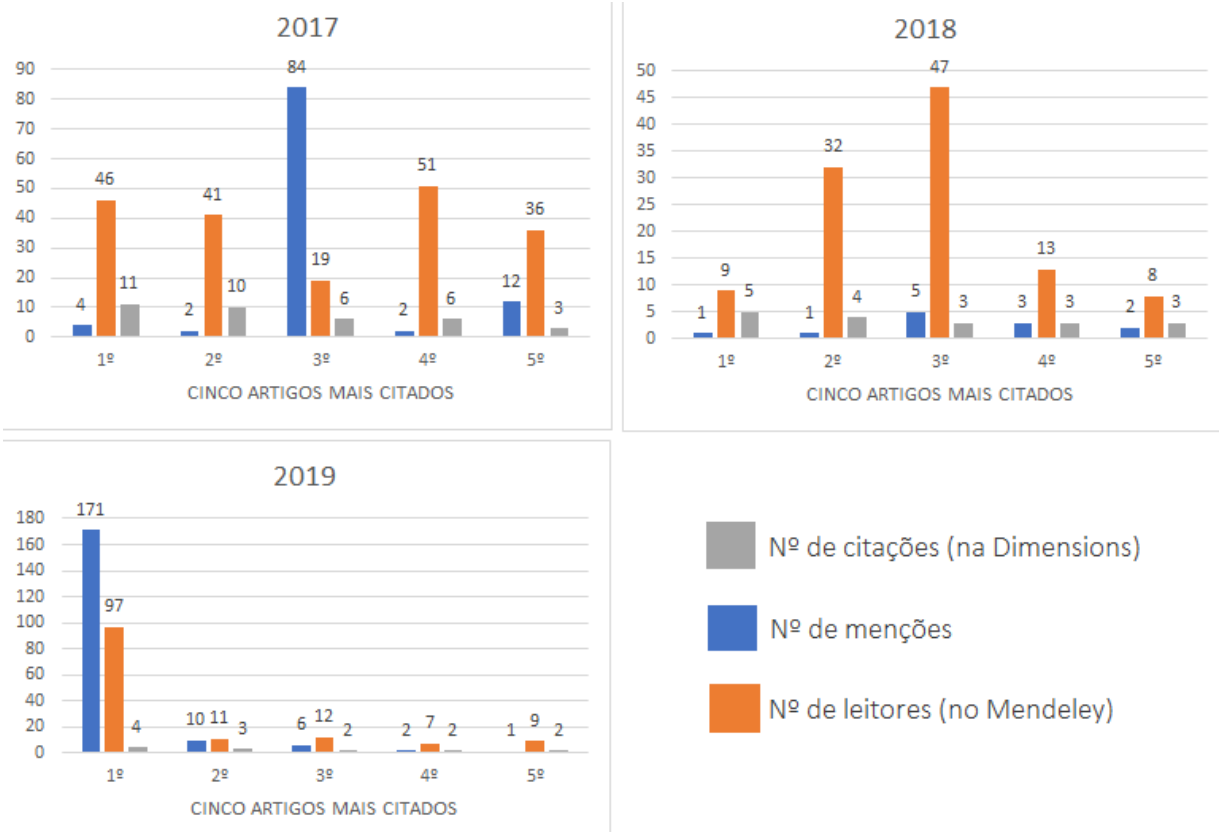
Os dados disponibilizados pela Altmetric possibilitam análises específicas e recortes específicos em razão dos objetivos desejados, tal como as análises temáticas dos artigos segundo áreas específicas de interesse. A partir dos dados do universo recuperado da coleção SciELO (81.622), foram analisados, a título de exemplo, 15 artigos das áreas da Ciência da Informação e da Computação.

Em seguida, os artigos foram ordenados pelo maior

número de citações (indexadas na Dimensions) e separados por data de publicação, considerando os três anos mais recentes (2017, 2018, 2019). Com os dados disponibilizados pela plataforma, é possível calcular correlações entre os anos, as áreas, as mídias e as citações. Por exemplo, a partir dos dados analisados constata-se que a correlação entre o número de leitores no Mendeley e as citações nos cinco artigos com maiores citações nos anos de 2017, 2018 e 2019, foram respectivamente: 0,3 (relação positiva fraca); -0,25 (relação negativa fraca); e, 0,9 (relação positiva muito forte).

Outras análises podem ser feitas a partir dos dados disponibilizados pela plataforma. Dessa amostra de 15 artigos, percebeu-se que os cinco artigos mais citados em cada ano foram mencionados em mídias sociais pelo menos uma vez e baixados por, no mínimo, sete leitores da biblioteca do Mendeley (figura 5).

Figura 5 – Relação entre menções e citações nos artigos das áreas da Ciência da Informação e da Computação (2017 - 2019)



Fonte: Dados da pesquisa (2019).

Observou-se também que o artigo com maior número de citações (11), publicado em 2017, não tem o maior número de menções e leitores. Assim como o artigo com maior número de menções (171) e leitores (97), publicado em 2019, não apresenta o maior número de citações. A partir desses dados, seria possível especular, por exemplo, que o número modesto de menções em 2017 deu-se em razão das mídias sociais serem menos utilizadas naquele ano. Ou que - como os indicadores altmétricos atingem o pico logo após a publicação e as citações levam tempo consideravelmente maior para serem acumuladas - o número reduzido de citações em 2019 deu-se ao fato do artigo ser recém-publicado.

A partir desse exercício, evidencia-se que os dados sobre menções e citações recuperados pela Altmetric viabilizam múltiplas aproximações. Seria possível, entre outras análises, fazer observações longitudinais, rastrear padrões e estabelecer diferentes correlações entre altmetria e contagem de citações. Apesar da quantidade e da qualidade dos dados apresentados, algumas limitações podem ser observadas. Alguns exemplos são a ausência de diversas afiliações institucionais dos autores (aproximadamente 25%) e a ausência de informações sobre língua do artigo e país da revista, impossibilitando diversos estudos.

Considerando as análises apresentadas na figura 5, além das outras limitações apresentadas, salienta-se que para fazer observações mais conclusivas são necessárias abordagens quantitativas e qualitativas capazes de identificar e questionar variâncias entre datas de publicação / citação / menção, plataformas de mídia social, tipos de menções, perfis de usuários e áreas de conhecimento. Embora os resultados deste trabalho e estudos semelhantes (HUANG; WANG; WU, 2018) indiquem que o número de citações pode ser influenciado pelo número de menções, essas inferências não podem surgir de uma correlação simplista de causa e efeito, que ignore a complexidade do fenômeno e a organicidade das redes.

CONSIDERAÇÕES FINAIS

Há décadas, pesquisadores da ciência da informação e gestores científicos rastreiam citações de publicações acadêmicas para medir, avaliar e compreender comportamentos nas ciências. Porém, é cada vez maior o número de estudos que concordam que tal abordagem ignora aspectos valiosos para análises da produção científica mais holísticas e robustas (PRIEM; HEMMINGER, 2010; HAUSTEIN, 2012; ADIE; ROE, 2013; NASSI-CALÒ, 2015, 2017). Diante da crise das métricas tradicionais e do uso crescente das mídias sociais para a disseminação das pesquisas, surgiram novas ferramentas para produzir indicadores sobre impactos da ciência na web 2.0, conhecidos como indicadores altmétricos. Uma das ferramentas mais utilizadas para este fim tem sido a Altmetric (ROBINSON-GARCIA *et. al.*, 2014).

Neste trabalho, analisou-se a Altmetric como plataforma fornecedora de dados e de indicadores aptos para compreender o impacto altmétrico de publicações científicas. Como foi observado, a plataforma Altmetric tem potencial para fornecer nova e ampla gama de dados sobre impactos da pesquisa, incluindo informações sobre audiências variadas (pesquisadores, profissionais, público em geral), números sobre o engajamento do público com as pesquisas (menções, compartilhamentos, downloads, etc) e números de citações feitas em resultados de pesquisas indexados pela Dimensions.

A partir da análise exploratória da base, constata-se que a Altmetric produz não apenas dados sobre as contagens e menções feitas à determinada pesquisa, mas também dados sobre usuários, autores da menção e da publicação, origem das menções, periódicos, status de *Open Access* e datas de cada publicação e menção. Dessa forma, a variedade de dados recuperados pela Altmetric abre outras possibilidades de análises adicionais, que vão além da simples contagem de menções. Por exemplo, a possibilidade de analisar tipos de público e participação por país na divulgação das pesquisas.

Entre as principais limitações identificadas estão: a falta de informações sobre a precisão e a exatidão das informações recuperadas pela base, especialmente para identificar menções em fontes mais complexas como blogs, sites de notícias e documentos de políticas públicas. Nestas mídias, o mecanismo de rastreamento está baseado em técnicas de mineração de texto, aplicadas como um complemento ao método de reconhecimento de link. No entanto, a base não deixa claro quais são os critérios seguidos na curadoria manual. Sabe-se apenas que a técnica aplicada compreende exclusivamente fontes de língua inglesa (um viés que deve ser fortemente considerado ao desenvolver estudos a partir dessas métricas). A este problema, somam-se casos em que os links estão indisponíveis (quebrados).

Outra limitação identificada está relacionada à escolha das mídias indexadas pela base. A partir da entrevista realizada com a diretora de relações de pesquisa, Stacy Konkiel, observou-se que a Altmetric "seleciona fontes com base em vários fatores, incluindo: quão amplamente a plataforma é usada, com que frequência a pesquisa é mencionada em uma fonte, se há demanda pelos dados e se é tecnicamente viável coletar dados daquele site" (KONKIEL, 2019). Mas, se o Weibo (site de língua chinesa) é uma das fontes selecionadas, por que não Tuenti espanhol? Seguindo essa linha, já que as métricas de leitores do Mendeley são computadas, por que não redes sociais científicas como a Academia.edu ou ResearchGate?

Ademais, os indicadores da Altmetric são derivados apenas de resultados de códigos e identificadores como PubMed ID, arXiv ID, ADS ID, SSRN ID, RePEC ID, Handle.net, URN, ISBNs e/ou DOIs recuperáveis. Apesar da amplitude, isso limita o conteúdo disponível essencialmente para aqueles dados identificáveis pelas ferramentas de harvesting da Altmetric. Contudo, como argumenta Konkiel (2019), os indicadores fornecidos pela Altmetric ajudam a identificar 'quem' e 'o que' está sendo falado sobre a pesquisa no ambiente online, possibilitando uma compreensão mais holística de impacto que envolve, além das citações, o alcance da pesquisa fora dos muros da academia.

Contatou-se ainda, que a Altmetric confunde e falha ao explicar qual é, exatamente, a metodologia e a base de cálculos utilizada no *Altmetric score*. Tal falta de clareza no procedimento impossibilita a auditoria do índice por parte da academia e a solução de problemas rotineiros do marketing científico, como saber, por exemplo, em que canais de divulgação as revistas deveriam investir mais. Também não foram encontradas na literatura sobre altmetria quaisquer evidências que explicitem como os cálculos deste índice são realizados. Investigar os mecanismos por trás de seu algoritmo é a pauta para pesquisas futuras.

Resguardados os problemas, com base nos dados levantados, conclui-se que a Altmetric é capaz de reunir dados altmétricos valiosos para análise da produção científica, a começar pela observação a nível individual dos resultados de pesquisa (artigos, ao invés de periódicos, por exemplo). Da mesma maneira que oportuniza diversas análises sobre 'como', 'onde' e 'por quem', a pesquisa tem sido percebida na web 2.0 e múltiplas aproximações entre impacto acadêmico (contagem de citações) e impacto social (contagem de menções).

Sobre as limitações deste estudo, admite-se que a perspectiva apontada não encerra a discussão sobre as possibilidades de pesquisas oferecidas pela Altmetric. Tampouco sobre as possibilidades de estudos que podem ser abordados pela altmetria. Uma vez que se compreende a volatilidade das trocas de informações geradas na Internet e a complexidade inerente à relação Ciência, Tecnologia e Sociedade, defende-se que o fenômeno da comunicação científica carece de todas as ferramentas disponíveis para análise e construção de novos indicadores. Além disso, admite-se que a realidade discutida aqui também não representa outras relevantes ferramentas altmétricas desenvolvidas até o momento (a saber, Impactstory, PlumX e Kudos, entre outras). Pesquisas futuras podem comparar as potencialidades entre as bases e oferecer luz sobre suas limitações e sobre as suas potencialidades e limites para a construção de indicadores altmétricos.

REFERÊNCIAS

ADIE, E.; ROE, W. Altmetric: enriching scholarly content with article-level discussion and metrics. *Learned Publishing*, [*s.l.*], v. 26, n. 1, p. 11-17, Jan. 2013. Disponível em: https://doi.org/10.1087/20130103. Acesso em: jan. 2021.

AHMED, W. Using Twitter as a data source: an overview of social media research tools. *LSE Impact Blog*. London, 8 May 2017. Disponível em: https://blogs.lse.ac.uk/impactofsocialsciences/2017/05/08/using-twitter-as-a-data-source-an-overview-of-social-media-research-tools-updated-for-2017/. Acesso em: 25 nov. 2019. Updated for 2017.

ALPERIN, J. P. Ask not what Altmetrics can do for you, but what altmetrics can do for developing countries. *Bulletin of the American Society for Information Science and Technology*, [*s.l.*], v. 39, n. 4, Apr. 2013. Disponível em: https://doi.org/10.1002/bult.2013.1720390407. Acesso em: 30 out. 2017.

ALTMETRIC. About our data. *Altmetric*, England, 2019. Disponível em: https://www.altmetric.com/about-our-data/. Acesso em: 2 dez. 2019.

BARROS, M. Altmetrics: métricas alternativas de impacto científico com base em redes sociais. *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 20, n. 2, jun. 2015. Disponível em: https://doi.org/10.1590/1981-5344/1782. Acesso em: jan. 2021.

CASTELLS, M. *A sociedade em rede.* São Paulo: Paz e Terra, 1999. (A era da informação: economia, sociedade e cultura).

COSTAS, R.; ZAHEDI, Z.; WOUTERS, P. Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology* [*s.l.*], v. 66, n. 10, p. 2003-2019, 2014. Disponível em: https://doi.org/10.1002/asi.23309. Acesso em: dez. 2019.

FAIRCLOUGH, R.; THELWALL, M. National research impact indicators from Mendeley readers. *Journal of Informetrics*, [*s.l.*], v. 9, n. 4, p. 845-859, 2015. Disponível em: https://doi.org/10.1016/j.joi.2015.08.003. Acesso em: dez. 2019.

HAUSTEIN, S. *Multidimensional journal evaluation*: analyzing scientific periodicals beyond the impact factor. Berlim: De Gruyter/Saur, 2012. (Knowledge & information).

HUANG, W.; WANG, P.; WU, Q. A correlation comparison between Altmetric Attention Scores and citations for six PLOS journals. *PLoS ONE*, [*s.l.*], v. 13, n. 4, 2018. Disponível em: https://doi.org/10.1371/journal.pone.0194962. Acesso em: 2 dez. 2019.

HUGHES, D. J. *et al*. A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior*, [*s.l.*], v. 28, n. 2, p. 561–569, 2012. DOI http://dx.doi.org/10.1016/j.chb.2011.11.001. Disponível em: http://opus.bath.ac.uk/28062/ . Acesso em: 22 out. 2019.

KONKIEL, S. A altmetria na Altmetric*:* uma entrevista com Stacy Konkiel. [Entrevista concedida a] Janinne Barcelos, Diego José Macedo, João de Melo Maricato. *RICI*, Brasília, 2019. No prelo.

LEMOS, A. *Cibercultura, Tecnologia e vida social na cultura contemporânea*. Porto Alegre: Sulina, 2002.

NASSI-CALÒ, L. A miopia dos indicadores bibliométricos. *SciELO em Perspectiva*, 1 jun. 2017. Disponível em: https://blog.SciELO.org/blog/2017/06/01/a-miopia-dos-indicadores-bibliometricos/#.Xe1Bv-hKjIU. Acesso em: 3 nov. 2019.

NASSI-CALÒ, L. Estudo analisa o uso de redes sociais na avaliação do impacto científico. *SciELO em Perspectiva*, 13 mar. 2015. Disponível em: http://blog.SciELO.org/blog/2015/03/13/estudo-analisa-o-uso-de-redes-sociais-na-avaliacao-do-impacto-cientifico/#.WYuns4jyvIU. Acesso em: 20 nov. 2019.

O'NEIL, J. NISO recommended practice: outputs of the Alternative Assessment Metrics Project. *Collaborative Librarianship*, [*s.l.*], v. 8, n. 3, p. 118–123, 2016. Disponível em: https://digitalcommons.du.edu/collaborativelibrarianship/vol8/iss3/4/ . Acesso em: 22 nov. 2019.

PIWOWAR, H. A. Altmetrics: Value all research products. *Nature*, [*s.l.*], v. 493, n. 159, 2013. Disponível em: https://www.nature.com/articles/493159a. Acesso em: 21 out. 2019.

PRIEM, J. *et al.* Altmetrics: a manifesto. [*S.l.*], 26 out. 2010. Disponível em: http://altmetrics.org/manifesto. Acesso em: 31 out. 2019.

PRIEM, J.; HEMMINGER, B. M. Scientometrics 2.0: toward new metrics of scholarly impact on the social Web. *First Monday*, Bridgman, v. 15, n. 7-5, 2010. Disponível em: https://firstmonday.org/article/view/2874/2570. Acesso em: 7 nov. 2019.

ROBINSON-GARCIA, N. *et al.* New data, new possibilities: exploring the insides of Altmetric.com. *El profesional de la información*, [*s.l.*], v. 23, n.4, p. 359-366, 2014. Disponível em: https://arxiv.org/abs/1408.0135. Acesso em: 15 nov. 2019.

ROEMER, R. C.; BORCHARDT, R. *Altmetrics*. Chicago: American Library Association, 2015.

TORRES-SALINAS, D.; CABEZAS-CLAVIJO, A.; JIMENEZ-CONTRERAS, E. Altmetrics: new indicators for scientific communication in web 2.0. *Comunicar*, Espanha, v. 21, n. 41, 2013. DOI 10.3916/C41-2013-05. Disponível: https://arxiv.org/abs/1306.6595. Acesso: 22 nov. 2019.

Estimando futuras colaborações com dados sobre atividades científicas

Thiago Magela Rodrigues Dias
Doutor em Modelagem Matemática e Computacional pela Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Belo Horizonte, Minas Gerais, Brasil. Professor do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) - Belo Horizonte, MG - Brasil.
http://lattes.cnpq.br/4687858846001290
E-mail: thiagomagela@gmail.com

Submetido em: 26/09/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021.

RESUMO

Em uma rede de colaboração científica, uma conexão é formada quando dois ou mais cientistas publicam um trabalho em conjunto. Nesse caso, as publicações representam as arestas e os cientistas, os nós da rede. Lançando mão de conceitos de análise de redes sociais, é possível compreender melhor o relacionamento entre os nós. O trabalho em questão tem o objetivo de realizar a predição de ligações em redes de coautoria formadas pelos doutores com currículos cadastrados na Plataforma Lattes, e que tenham, como área de atuação, as Ciências da Informação. Atualmente, a Plataforma Lattes conta com 6.6 milhões de currículos de indivíduos e representa um dos conjuntos de dados curriculares mais relevantes e reconhecidos mundialmente. Diante isso, é possível compreender o comportamento da rede e acompanhar a sua evolução ao longo do tempo. Para tanto, algumas etapas precisam ser seguidas. São elas: extração dos dados, criação das redes de coautoria, definição dos atributos a serem utilizados, criação de um conjunto de dados, e por fim, emprego dos mesmos como entrada em um algoritmo de aprendizado de máquinas. Por meio dos resultados, é possível estabelecer, com precisão, a evolução da rede de colaborações científicas dos pesquisadores a nível nacional, auxiliando, assim, as agências de fomento na escolha de futuros pesquisadores de destaque.

Palavras-chave: Colaboração científica. Predição de ligações. Plataforma Lattes.

Estimating future collaborations with data on scientific activities

ABSTRACT

In a scientific collaboration network, a connection is formed when two or more scientists publish a work together. In this case, the publications represent the edges, and the scientists represent the nodes of the network. Using concepts from the analysis of social networks, it is possible to better understand the relationship between nodes. The work in question aims to make the prediction of connections in co-authorship networks formed by PhDs with curricula registered in the Lattes Platform, and whose area of activity is Information Sciences. Currently, the Lattes Platform has 6.6 million curricula of individuals and represents one of the most relevant and recognized scientific repositories worldwide. With this, it is possible to understand the behavior of the network and monitor its evolution over time. For that, some steps are necessary, they are: data extraction, creation of co-authorship networks, definition of the attributes to be used, creation of a data set, and finally, use them as input in a machine learning algorithm. Through the results it is possible to establish, with precision, the evolution of the network of scientific collaborations of the researchers at national level, thus assisting the funding agencies in the choice of future outstanding researchers.

Keywords: *Scientific collaboration. Link prediction. Lattes Plataform.*

Estimación de colaboraciones futuras con datos sobre actividades científicas

RESUMEN

En una red de colaboración científica, se forma una conexión cuando dos o más científicos publican un trabajo en conjunto. En tal caso, las publicaciones representan los bordes y los científicos representan los nodos de la red. Aprovechando los conceptos de análisis de redes sociales, es posible comprender mejor la relación entre nodos. El trabajo en cuestión tiene como objetivo hacer la predicción de las conexiones en redes de coautoría formadas por médicos con currículos registrados en la Plataforma Lattes, y cuya área de actividad son las Ciencias de la Información. Actualmente, la Plataforma Lattes tiene 6.6 millones de CV de personas y representa uno de los repositorios científicos más relevantes y reconocidos a nivel mundial. Con esto, es posible comprender el comportamiento de la red y monitorear su evolución a lo largo del tiempo. Para eso, algunos pasos son necesarios, son: extracción de datos, creación de redes de coautoría, definición de los atributos que se utilizarán, creación de un conjunto de datos y, finalmente, utilizarlos como entrada en un algoritmo de aprendizaje de máquinas. A través de los resultados es posible establecer, con precisión, la evolución de la red de colaboraciones científicas de los investigadores a nivel nacional, ayudando así a las agencias de financiación en la elección de futuros investigadores destacados.

Palavras clave: *Colaboración científica. Predicción de enlaces. Plataforma Lattes.*

INTRODUÇÃO

No final da década de 90, diversos pesquisadores dedicaram atenção aos estudos de redes. Foram realizados trabalhos sobre a área da biologia, da internet, de roteadores, entre outros (NEWMAN, 2001; NEWMAN; PARK, 2003; BARABÁSI; ALBERT, 1999). Tais investigações permitiram entender o relacionamento entre os nós, fazendo surgir, ao se estudar essas ligações, a pergunta: "como ocorre a evolução da rede ao longo do tempo?". Hasan e Zaki (2011), porém, explicam que compreender a evolução da rede como um todo é uma tarefa complexa.

Com esses conceitos em mente, Liben-Nowell e Kleinberg (2003) propuseram o problema da predição de ligações. Inicialmente, foram utilizados métodos que calculavam a similaridade entre dois nós da rede. Quanto mais parecidos, maior a chance de possuírem uma ligação entre si. A partir de então, diversos outros métodos foram propostos para melhor resolução do problema da predição de ligações (ACAR; DUNLAVY; KOLDA, 2009; ZHOU; LÜ; ZHANG, 2009; LIU *et al.*, 2011).

Atualmente, são empregados métodos probabilísticos, métodos baseados em álgebra linear e, também, métodos que transformam esse problema em um de classificação binária, dessa forma, diversos algoritmos podem ser aplicados à sua resolução. Neste artigo, tratamos a predição de ligações como um problema de classificação, assim, algoritmos da área de sistemas de recomendação são empregados na realização dos objetivos propostos.

Aplicando tais conceitos a um domínio mais específico, podemos dirigir as atenções às redes pertencentes à comunidade científica. Ao se publicar uma comunicação científica com outro cientista, uma ligação é formada pela colaboração realizada. Nessas redes, os autores representam os nós e as colaborações científicas, as arestas (MARUYAMA; DIGIAMPIETRI, 2019). Tais redes são chamadas de redes de coautoria e serão o objeto de estudo deste artigo.

Nesse contexto, a Plataforma Lattes, mantida pelo CNPQ[1], tem sido fonte de dados de diversos trabalhos que visam a analisar redes de colaboração científica, principalmente por englobar dados de grande parte da produção científica nacional.

Cañibano e Bozeman (2009) destacam que os currículos acadêmicos são fontes de informação potenciais e extremamente abrangentes, bem como foco de investigações recentes que estudam grupos de pesquisadores. Inquirições que se valem de currículos no exame de redes sociais são ainda menos frequentes, porém, deve-se considerar a gama de trabalhos sobre análise de coautoria e os efeitos das colaborações científicas na carreira do pesquisador (DIGIAMPIETRI; SANTIAGO; ALVES, 2013; LIMA *et al.*, 2013; MENA-CHALCO; CESAR-JUNIOR, 2013).

Perez-Cervantes *et al.* (2013) introduzem novas medidas para estimar a influência da colaboração em redes científicas. A abordagem é baseada na técnica de predição de *links* e avalia como a presença ou ausência de um pesquisador afeta o processo de predição na rede em exame. Para isso, os cientistas são representados por nós em uma rede de colaboração e, após a remoção de nós, o processo de predição de *links* é realizado de forma iterativa para todos os outros nós.

Já Mena-Chalco *et al.* (2014) utilizam dados dos currículos da Plataforma Lattes para identificar e caracterizar a rede de colaboração de pesquisadores brasileiros. Essa pesquisa objetiva extrair os dados de currículos cadastrados na Plataforma Lattes, identificar automaticamente a colaboração baseada em informações bibliométricas, produzindo uma rede de colaboração, e aplicar métricas baseadas em análise topológica para compreender como ocorre a interação entre os pesquisadores.

Por sua vez, Sidone, Haddad e Mena-Chalco (2016) apresentam o papel da geografia na evolução da produção e colaboração científica no Brasil entre 1992 e 2009.

Nesses estudos, foi feito uso de dados dos currículos de um milhão de pesquisadores, abrigados na Plataforma Lattes. Os autores destacam o processo de desaceleração da produção científica brasileira a partir dos últimos triênios analisados.

Atualmente, a Plataforma Lattes conta com 6.6 milhões de currículos cadastrados e representa uma das mais relevantes fontes de dados sobre atividades científicas e pesquisadores, além de ser reconhecida mundialmente (LANE, 2010). O conjunto de dados registrados nos currículos cadastrados nesse sistema de informações possui atributos como: nome, formação acadêmica, experiência profissional, projetos, publicações científicas, entre outros. O grande volume de dados presente nos currículos pode fornecer informações valiosas e, até então, desconhecidas (DIAS *et al.*, 2013).

Dessa forma, será realizada a predição de ligações em redes de coautoria, formada pelos dados de doutores presentes em currículos cadastrados na Plataforma Lattes. Com isso, será possível compreender o comportamento dessa rede e acompanhar a sua evolução ao longo do tempo. Por meio deste estudo, também será possível identificar os pesquisadores que poderão colaborar com a rede no futuro.

METODOLOGIA

Para que seja possível atingir os objetivos propostos, é essencial que se siga alguns passos. Desse modo, nesta seção, serão destacados os métodos empregados nesta pesquisa para que seja possível realizar a predição de futuras ligações em uma área específica. Para tanto, foi escolhida a grande área de Ciências Sociais Aplicadas e, posteriormente, a área de Ciência da Informação. Esse conjunto de dados possui 1.084 pesquisadores com título de doutor. Inicialmente, será apresentado o *framework* empregado na extração dos dados. Em um segundo momento, serão mostradas as redes de colaboração científicas criadas, e por último, os atributos selecionados para a predição serão caracterizados.

[1] Conselho Nacional de Desenvolvimento Científico e Tecnológico

Para início do desenvolvimento do trabalho, foi necessário extrair os dados a serem utilizados. Sendo assim, lançou-se mão de um *framework* para extração e tratamento dos dados, o *LattesDataExplorer* (DIAS, 2016). Após a coleta dos dados, ocorrida em 2019, as informações foram organizadas e, posteriormente, as redes foram caracterizadas, conforme método para identificação de colaborações científicas em grandes bases de dados, com uso de baixo poder computacional, apresentado por Dias e Moita (2015).

Após a caracterização das redes de colaboração, foi preciso identificar os atributos utilizados na predição. Assim, um conjunto básico de características, oriundo de outros trabalhos que abordaram esse tema, foram selecionados. Ainda tendo a Plataforma Lattes como fonte de dados, foi possível obter informações referentes ao domínio que estava sendo analisado, visto que os autores possuem registro de algumas informações pessoais, como a cidade e o estado em que residem, e a universidade da qual fazem parte. Como tais campos podem ser empregados no auxílio à predição a ser realizada, dois tipos de atributos são definidos: os atributos topológicos e os atributos referentes ao domínio, conforme demonstrado no quadro 1.

Enquanto atributos topológicos são obtidos a partir de alguns cálculos, que utilizam como base a própria rede, atributos referentes ao domínio são extraídos e armazenados da mesma forma que o autor preencheu as informações de seu currículo. Porém, ao se lançar mão de técnicas de aprendizado de máquinas, é importante que os dados estejam padronizados para facilitar o processo de predição. Os campos "cidade", "estado" e "instituição" são considerados dados categóricos, e, por isso, devem passar por duas etapas antes de serem aplicados ao restante do processo.

Inicialmente, é necessário codificar os textos informados pelo autor em números. Por exemplo: no lugar de "Belo Horizonte", o valor 5 será armazenado; para "São Paulo", o valor 13; e assim por diante, para todas as informações categóricas. Para fazer essa codificação, é feito uso do método *Label Encoding*, algoritmo que realiza o passo a passo descrito acima para os dados selecionados. Dessa forma, todos os valores categóricos são codificados em números. Porém, após esse processo, os algoritmos a serem utilizados, podem identificar, a título de ilustração, que o valor 13, referente a São Paulo, é mais vultoso do que o valor 5, referente a Belo Horizonte, afinal, não foi especificado que esses valores representam categorias. Para evitar que isso aconteça, outro método, o *One Hot Encoder,* deve ser aplicado. Por meio dele, cada categoria é transformada em uma coluna, e, caso o valor seja referente a uma determinada coluna, é inserido o número 1, caso contrário, é inserido o zero. Assim, os dados categóricos são convertidos em uma grande matriz esparsa, composta, em sua maioria, pelo número zero.

Após a definição dos atributos, alguns passos devem ser seguidos. Em um primeiro momento, é necessário definir os períodos para treino e teste, de modo que três redes diferentes foram caracterizadas. Para a rede 1, foram definidas as publicações realizadas no período entre 1960 e 2000, que será chamado de período inicial. Já a segunda rede foi caracterizada pelo período de 2001 a 2010. Por fim, foi estabelecido o período de 2011 a 2018 para a terceira e última rede. Tais períodos compreendem a data do primeiro trabalho registrado na plataforma até o último ano anterior à coleta dos dados.

As informações referentes às redes são exibidas na tabela 1 e na figura 1, onde é possível perceber as mudanças ao longo dos períodos analisados. No princípio, nota-se a pequena quantidade de colaborações, que impacta diretamente no grau médio da rede, e na sua densidade.

Quadro 1 – Definição de atributos utilizados no processo de predição

Atributo	Descrição	Tipo
Vizinhos em Comum (VC)	De acordo com Liben-Nowell e Kleinberg (2003), a forma mais simples de realizar a predição de arestas é por intermédio da métrica Vizinhos em Comum, que pode ser entendida como a quantidade de nós em comum que dois nós específicos possuem.	Topológico
Coeficiente de Jaccard (JC)	Mede a probabilidade de que ambos, x e y, possuam um vizinho v, escolhido aleatoriamente. Hasan e Zaki (2011) explicam que, ao contrário do atributo Vizinhos em Comum, o coeficiente de Jaccard normaliza o número de vizinhos em comum.	Topológico
Adamic/Adar (AA)	Essa formulação atribui, às características mais raras, um peso maior. Podemos entendê-la como o número de propriedades compartilhadas pelos nós, dividido pelo log da frequência das características.	Topológico
Resource Allocation (RA)	Seguindo o mesmo raciocínio, a métrica Resource Allocation atribui peso na relação de dois nós, favorecendo as relações entre aqueles que possuem poucos relacionamentos.	Topológico
Preferential Attachment (PA)	A métrica Preferential Attachment foi proposta considerando apenas o tamanho das vizinhanças dos nós,. Em suma, ela estabelece que a probabilidade de um novo relacionamento com outros vértices é baseada no grau do nó em questão.	Topológico
Menor Caminho (MC)	O fato de que amigos de amigos podem criar uma ligação sugere que a distância entre os nós de uma rede pode influenciar na formação de novas ligações. Podemos entendê-la como o caminho mínimo entre dois nós.	Topológico
Colaborações em conjunto (Peso)	Dessa forma, é possível identificar colaboradores que já trabalham juntos há mais tempo e, possivelmente, possuem uma maior influência nos próximos instantes de tempo.	Topológico
Instituição	Instituição à qual o pesquisador está vinculado.	Domínio
Estado	Estado cadastrado no campo “endereço profissional”, no do currículo do pesquisador.	Domínio
Cidade	Cidade cadastrada no campo “endereço profissional”, no currículo do pesquisador.	Domínio

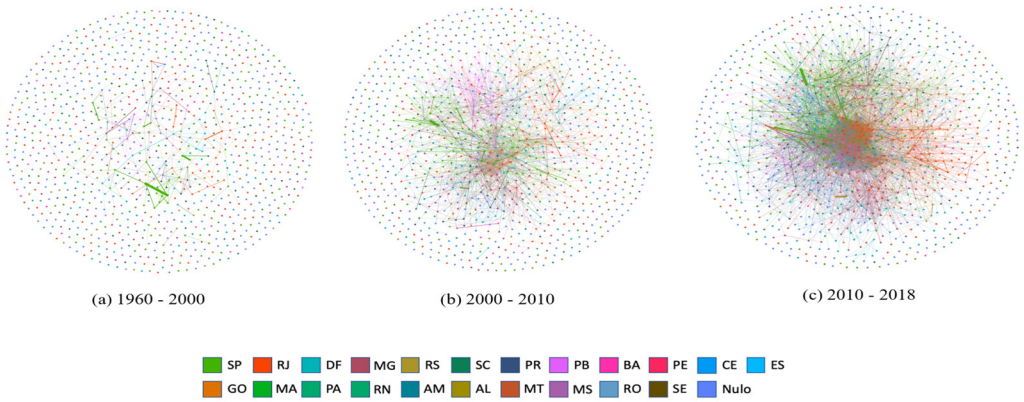
Fonte: Adaptado de Adamic, L. A. e Adar, E. (2003); Chen, H., Li, X., Huang, Z. (2005); Liben-Nowell, D. e Kleinberg, J. (2007); Potgieter, A. *et al.* (2009); Digiampietri, L. *et al.* (2015); Maruyama, W. T. e Digiampietri, L. A. (2019); Hasan, M. A. e Zaki, M. J. (2011); Lü, L. e Zhou, T. (2011).

Tabela 1 – Redes de colaboração científica caracterizadas para o estudo

Período	Pesquisadores	Colaborações	Grau médio	Densidade	Diâmetro	Caminho médio
1960 – 2000	1.084	1.064	0,466	0	15	5,523
2000 – 2010	1.084	6.186	3,701	0,003	10	4,081
2010 – 2018	1.084	11.603	11,051	0,01	8	3,280

Fonte: Elaboração dos autores.

Figura 1 – Visão geral das redes de colaboração científica caracterizadas.



Fonte: Dados da Pesquisa.

É importante examinar o aumento da densidade no decurso do tempo, uma vez que esse fator influencia diretamente no problema chamado desbalanceamento de classes, que será apresentado posteriormente. Pode-se entender o grau médio, nesse caso, como a média das colaborações realizadas pelos pesquisadores. Ao final do período analisado, o número médio de colaborações aumentou consideravelmente, refletindo, assim, na densidade da rede. Outro fator que indica a evolução da rede é a diminuição no caminho médio, demonstrando que os pesquisadores criaram mais conexões entre si, ficando, dessa maneira, mais fácil de se conectar a novos parceiros de pesquisa. Também cabe destacar que, para os propósitos desta investigação, o número de nós se manteve constante durante todos os períodos.

O conjunto de dados contendo os pesquisadores, as ligações entre si e os atributos selecionados foi então utilizado como entrada em um algoritmo de aprendizado de máquina.

Cada linha do conjunto de dados é composta pelos seguintes itens: identificação do primeiro pesquisador, identificação do segundo pesquisador, vizinhos em comum, coeficiente de Jaccard, Adamic/Adar, *Resource Allocation*, *Preferential Attachment*, Menor Caminho, peso, presença (ou ausência) de uma aresta, instituição, cidade e estado.

Nessa etapa do trabalho, o problema do desbalanceamento de classes vem à tona. O número de ligações possíveis em um grafo é quadraticamente relacionado ao número de nós, no entanto, o cômputo de ligações reais representa apenas uma pequena fração desse número (HASAN; ZAKI, 2011).

Uma técnica tradicional para superar o desbalanceamento das classes é chamada de sob amostragem. Ela consiste em reduzir o número de amostras da classe determinante, de forma randômica, igualando, assim, o número de componentes para ambos os casos. Essa técnica foi aplicada à investigação aqui apresentada.

No início, o conjunto de dados apresentava uma proporção de 152 arestas ausentes para cada aresta presente. Após a aplicação da sob amostragem, o número de arestas presentes e ausentes foi o mesmo. Com os dados balanceados, o algoritmo para predição de ligações foi executado.

RESULTADOS

Ao longo do processo descrito na seção anterior, o conjunto de dados sofreu algumas alterações. No total, os 1.084 pesquisadores podem possuir um total de 586.896 arestas. Destas, apenas 3.831 representavam arestas positivas na Rede 3. Logo, por meio do balanceamento das amostras, um conjunto randômico de outras 3.831 arestas ausentes foi escolhido. Sendo assim, o conjunto de dados utilizado na entrada no algoritmo de predição de dados é composto por 7.662 registros. Ao se fazer uso dos métodos de aprendizado de máquinas, é importante separar uma parte do conjunto de informações para treinar o algoritmo, e outra parte para o teste do mesmo. Esse segundo conjunto deve possuir dados até então não empregados em algum momento pelos algoritmos de predição, de modo a validar que realmente ocorreu um aprendizado, e não apenas um condicionamento dos valores já utilizados. Dessa forma, foram selecionadas 5.746 ligações (escolhidas aleatoriamente) para treino, representando 25% do conjunto total, e outras 1.916 ligações para teste.

Diversos algoritmos podem ser aplicados à resolução de problemas de classificação. Entre eles, alguns foram selecionados para execução do trabalho, tais como: Regressão Logística, K-Vizinhos Mais Próximos, Baías Ingênuas e Florestas Aleatórias.

Cada uma dessas técnicas possui uma particularidade diferente e, por conseguinte, consequências distintas. Portanto, seus resultados serão evidenciados na tabela 2, lançando mão das métricas: precisão, revocação, F1 e área sob a curva (AUC). Como, normalmente, a maioria dos autores faz uso da área sob a curva em algoritmos empregados na predição de ligações, ela também é utilizada como base nesta análise.

Tabela 2 – Resultados utilizando atributos topológicos da rede

Algoritmo	Precisão	Revocação	F1	AUC
Regressão Logística	0.67	0.66	0.65	0.70
K-Vizinhos Mais Próximos	0.71	0.68	0.68	0.71
Baías Ingênuas	0.76	0.62	0.56	0.70
Florestas Aleatórias	0.70	0.68	0.67	0.71

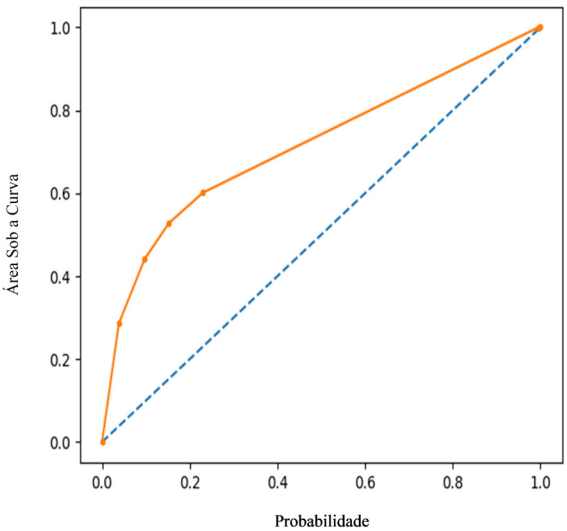
Fonte: Dados da Pesquisa, 2019.

Cada uma das métricas aplicadas à validação dos resultados possui características próprias. A precisão tem como objetivo responder à seguinte pergunta:De todos os valores preditos positivos, quantos realmente estão corretos? Uma alta precisão está relacionada a poucos falsos positivos. Considerando todos os valores positivos, a revocação tem o objetivo de saber quantos destes foram realmente preditos. A métrica F1 leva em conta a precisão e a revocação, fazendo, assim, uma média ponderada dessas duas métricas. Por último, a área sob a curva, ou *area under the curve* (AUC), em inglês, é utilizada para exibir o desempenho de um modelo de classificação no decurso de todo o processo de aprendizagem.

Foram realizados dois processos de predição: o primeiro deles, servindo-se apenas dos atributos topológicos da rede; e, posteriormente, de todo o conjunto de dados, contendo os dois tipos de atributos: topológicos e relacionados ao domínio. Dessa forma, também, é possível examinar a importância de se estudar o contexto no qual a predição de ligações será realizada.

Analisando a tabela 2, que comporta os dados referentes ao processo de predição, fazendo uso apenas dos atributos topológicos, é possível perceber que os algoritmos escolhidos obtiveram bons resultados. Dessa forma, fica claro que o algoritmo conseguiu empregar o conjunto de dados e características ora apresentado para realizar predições corretas a respeito de futuras ligações. Ao observar a área sob a curva, percebemos que todos obtiveram um resultado acima do que um mero acaso. Essa situação é visualizada com mais nitidez na figura 2, onde a linha pontilhada em azul representa uma chance de 50% de acerto, ou seja, probabilidades iguais para a predição ser da classe correta ou incorreta, e a linha laranja representa os valores das predições realizadas.

Figura 2 – Área sob a curva (AUC) para o algoritmo K-Vizinhos Mais Próximos.



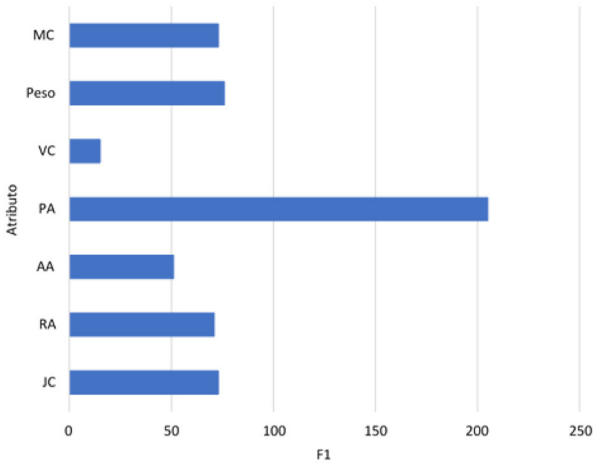
Fonte: Dados da Pesquisa, 2019.

Entre os algoritmos utilizados, o que demonstrou melhor desempenho, levando em conta todas as métricas, foi o K-Vizinhos Mais Próximos, seguido por Florestas Aleatórias, Baías Ingênuas, e, por último, Regressão Logística. Porém, existe uma pequena diferença entre os resultados obtidos, deixando claro que, para o problema em questão, ainda não podemos estabelecer qual técnica deveria ser tomada como padrão.

Ao se analisar o processo de aprendizado com base nos atributos utilizados, é possível identificar a ordem de influência de cada um deles no resultado final. Podemos observar, a partir da figura 3, que a ordem de importância dos atributos para a predição aqui realizada é: *Preferential Attachment*, Peso das Colaborações, Caminho Mais Curto, Coeficiente de Jaccard, *Resource Allocation*, Adamic/Adar, e, por fim, Vizinhos em Comum.

Tal fato apresenta um comportamento até então diferente da maioria dos referenciais teóricos aqui estudados, em que, na maior parte das vezes, o atributo mais relevante é o Vizinhos em Comum. Já, nos estudos aqui realizados, a métrica *Preferential Attachment* é responsável por boa parte do resultado final.

Figura 3 − Ordem de importância dos atributos para a predição



Fonte: Dados da Pesquisa, 2019.

Tabela 3 − Resultados considerando todos os atributos da rede

Algaritmo	Precisão	Revogação	F1	AUC
Regressão Logística	0.78	0.78	0.78	0.87
K-Vizinhos Mais Próximos	0.74	0.73	0.73	0.80
Baías Ingênuas	0.77	0.63	0.63	0.63
Florestas Aleatórias	0.77	0.77	0.77	0.85

Fonte: Dados da Pesquisa, 2019.

Ao olhar com atenção para toda a base de dados, contendo todos os atributos, os resultados revelaram uma expressiva melhora, conforme tabela 3.

Em média, as predições foram 8,25% melhores, sendo, a melhor técnica, a Regressão Logística, que conseguiu identificar futuras colaborações com 87% de certeza. Em segundo lugar, o algoritmo Florestas Aleatórias apresentou um resultado bem próximo, com 85% de acerto, seguido pelos K-Vizinhos Mais Próximos, com 80% de acerto, e, finalmente, as Baías Ingênuas.

Nesse segundo momento, os atributos não foram analisados separadamente, visto que os dados categóricos foram codificados em diversas colunas, transformando, assim, o conjunto de dados em uma matriz esparsa. Todas as outras métricas também apresentaram um resultado melhor do que apenas utilizando os atributos topológicos da rede.

CONSIDERAÇÕES FINAIS

Os resultados aqui expostos demonstram que é possível realizar a predição de ligações lançando mão de informações da própria rede estudada. O objetivo proposto foi então alcançado, uma vez que, a partir da utilização desses dados, é possível saber, por exemplo, se dois pesquisadores da área citada acima irão colaborar em um futuro instante de tempo. Um dos pontos mais importantes desta investigação está relacionado com a evolução da rede de colaboração científica. Com o passar do tempo, as colaborações saíram de uma média de 0,46 para 11,05 por pesquisador, demonstrando que o trabalho em equipe é cada vez mais necessário.

Observando os resultados obtidos com base nas predições, fica clara a importância de se possuir conhecimento sobre o domínio a ser analisado. Inicialmente, fazendo uso apenas de atributos topológicos, ou seja, referentes à própria rede, a melhor taxa de acertos foi de 71%, para os algoritmos K-Vizinhos Mais Próximos e Florestas Aleatórias, onde o atributo mais relevante foi o *Preferential Attachment*. Esse atributo demonstra que a probabilidade de um autor publicar uma comunicação científica varia conforme o número de colaborações já realizadas.

Considerando também os atributos referentes ao domínio, nesse caso, a instituição, a cidade e o estado do pesquisador, o aumento na quantidade de predições corretas foi expressivo, saltando de 70,5%, em média, para 78,75%. Desse modo, o algoritmo com o melhor resultado foi a Regressão Logística, que realizou a predição correta em 86% dos casos.

Os atributos categóricos utilizados para que houvesse essa melhora passaram por um longo processo de codificação, a fim de que tais resultados fossem alcançados. Dessa forma, é evidente o mérito do emprego das técnicas mais avançadas de aprendizado de máquinas, para que seja possível aumentar o número de predições corretas.

Em trabalhos futuros, destaca-se a importância de aumentar o conjunto de dados ou, até mesmo, buscar outras formas de solucionar o problema do desbalanceamento de classes, aumentando, assim, o número de amostras presentes para treino do algoritmo. Desse ponto em diante, espera-se que os classificadores apresentem um desempenho ainda melhor.

REFERÊNCIAS

ACAR, E.; DUNLAVY, D. M.; KOLDA, T. G. Link prediction on evolving data using matrix and tensor factorizations. In: *Proceedings of the workshop on large-scale data mining:* theory and applications (LDMTA'09). [s.l]: [s.n], 2009. p. 262-269.

ADAMIC, L. A.; ADAR, E. Friends and neighbors on the web. *Social Networks, Elsevier*, v. 25, n. 3, p. 211-230, 2003.

BARABÁSI, A. L. E.; ALBERT, R. Emergence of scaling in random networks. *Science, American Association for the Advancement of Science*, v. 286, n. 5439, p. 509-512, 1999.

CAÑIBANO, C.; BOZEMAN, B. Curriculum vitae method in science policy and research evaluation: the state-of-the-art. *Research Evaluation*, v. 18, n. 2, p. 86-94, 2009.

CHEN, H.; LI, X.; HUANG, Z. Link prediction approach to collaborative filtering. *Proceedings Of The ACM/IEE Joint Conference on Digital Libraries*, p. 141-142, 2005.

DIAS, T. M. *et al.* Modelagem e caracterização de redes científicas: um estudo sobre a Plataforma Lattes. Brasnam-Ii Brazilian Workshop On Social Network Analysis And Mining, p. 10-20, 2013.

DIAS, T. M. R.; MOITA, G. F. A method for the identification of collaboration in large scientific databases. *Em Questão*, v. 21, n. 2, p. 140-161, 2015.

DIAS, T. *Um estudo da produção científica brasileira a partir de dados da Plataforma Lattes*. Tese (Doutorado em Modelagem Matemática e Computacional) - Centro Federal de Educação Tecnológica de Minas Gerais. Belo Horizonte, 181 p., 2016.

DIGIAMPIETRI, L. A; SANTIAGO, C. R. N.; ALVES, C. M. Predição de coautorias em redes sociais acadêmicas: um estudo exploratório em Ciência da Computação. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 2, 2013, *Anais…* Maceió, 2013.

DIGIAMPIETRI, L. *et al.* Um sistema de predição de relacionamentos em redes sociais. *Brazilian Symposium on Information Systems*, v. 11, 2015.

HASAN, M. A.; ZAKI, M. J. A survey of link prediction in social networks. In: AGGARWAL, C. (Ed.). *Social network data analytics*. Boston: Springer, 2011, p. 243–275.

LANE, J. Let's Make Science Metrics More Scientific. *Nature*, v. 464, n. 7288, p. 488-489, 2010.

LIBEN-NOWELL, D.; KLEINBERG, J. The link-prediction problem for social networks. *Journal of The American Society For Information Science And Technology*, v. 58, n. 7, p. 1019-1031, 2007.

LIMA, H. *et al.* Aggregating productivity indices for ranking researchers across multiple areas. In: Proceedings of the 13th acm/ieee-cs joint conference on digital libraries, ACM, p. 97-106, 2013.

LIU, Z. *et al.* Link prediction in complex networks: a local naïve bayes model. *EPL (Europhysics Letters)*, v. 96, n. 4, 2011.

LÜ, L.; ZHOU, T. Link prediction in complex networks: a survey. *Elsevier*, v. 390, n. 6, p. 1150-1170.

MARUYAMA, W. T.; DIGIAMPIETRI, L. A. Co-Authorship Prediction In Academic Social Network. In: Workshop Brasileiro de Análise de Redes Sociais e Mineiração, V., 2019, Porto Alegre. *Anais…* Porto Alegre: Sociedade Brasileira de Computação, [2016].

MENA-CHALCO, J. P.; CESAR-JUNIOR, R. M. Prospecção de dados acadêmicos de currículos Lattes através de scriptLattes. In: HAYASHI, M. C. P. I.; LETA, H. E. J. (Orgs.). *Bibliometria e Cientometria*: reflexões teóricas e interfaces. São Carlos: Pedro & João, p. 109-128, 2013.

MENA-CHALCO, J. P.; *et al.* Brazilian bibliometric coauthorship networks. *Journal of the Association for Information Science and Technology*, v. 65, n. 7, p. 1424-1445, 2014.

NEWMAN, M. E. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences,* v. 98, n. 2, p. 404-409, 2001.

NEWMAN, M. E.; PARK, J. Why social networks are different from other types of networks. *Physical Review E.*, v. 68, n. 3, 2003.

PEREZ-CERVANTES, E. *et al.* Using link prediction to estimate the collaborative influence of researchers. In: IEEE 9TH INTERNATIONAL CONFERENCE ON ESCIENCE (ESCIENCE), IX. *Anais…* China, Beijin, p. 293-300, 2013.

POTGIETER, A. *et al.* Temporality in Link Prediction: Understanding Social Complexity. *Emergence, Complexity & Organization*, v.11, n.1, p.69-83, 2009.

SIDONE, O. J. G.; HADDAD, E. A.; MENA-CHALCO, J. P. A Ciência nas Regiões Brasileiras: Evolução da Produção e das Redes de Colaboração Científica. *Transinformação,* v. 28, n. 1, p. 15-31, 2016.

ZHOU, T., LÜ, L., ZHANG, Y.-C. Predicting Missing Links Via Local Information. *The European Physical Journal*, v.71, n.4, p. 623–630, 2009.